Sushil Tripathi

Laying the foundations for Gastrin Systems Biology

Conceptual models and knowledge resources to enhance research on gastrin mediated intracellular signaling and gene regulation

Thesis for the degree of Philosophiae Doctor

Trondheim, August 2013

Norwegian University of Science and Technology
Faculty of Medicine
Department of Cancer Research and Molecular Medicine

NTNU – Trondheim
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the degree of Philosophiae Doctor

Faculty of Medicine
Department of Cancer Research and
Molecular Medicine

ISBN 978-82-471-4562-3 (printed ver.)
ISBN 978-82-471-4563-0 (electronic ver.)
ISSN 1503-8181

Doctoral theses at NTNU, 2013:222

Printed by NTNU-trykk

Sammendrag

Behovet for å forstå biologisk kompleksitet har vært en sterk pådriver for utvikling av ny teknologi og nye kunnskapsressurser. Avanserte og kvantitative storskalateknologier har gjort det nærmest trivielt å framstille massive datamengder for et gitt biologisk system. I arbeidet med å generere kunnskap fra disse datamengdene, har avansert genomskala dataanalyse blitt en flaskehals. God forvaltning av eksisterende kunnskap som skapes av det globale vitenskapssamfunnet er sentralt for å kunne ekstrahere biologisk innsikt fra nye storskaladata.

I den foreliggende doktorgradsavhandlingen presenteres arbeid som bidrar til et grunnlag for systemnivå-analyse av gastrin-induserte signaltransduksjon-kaskader og cellulære responser ('gastrin-signalsystemet'). Arbeidet beskriver integrasjon av kunnskap fra mange ulike kilder som bidrar til en nettverksmodell for gastrin-signalsystemet. Denne nettverksmodellen kan spille en rolle for framtidig forskning og derigjennom bedre vår forståelse av gastrin-indusert signaltransduksjon og genreguleringsnettverk.

Transkripsjonsfaktorer utgjør et viktig bindeledd mellom signaltransduksjonskaskadene og påfølgende endringer i genuttrykk. For å bidra til presis og tilgjengelig kunnskap om transkripsjonsfaktorer (TFs), fokuserer avhandlingen dels på etablering av robuste retningslinjer for kuratering (målrettet registrering og sammenstilling av kunnskap) av transkripsjonsfaktorer som er eksperimentelt bekreftet og dokumentert i litteraturen, og dels på konstruksjon av TFcheckpoint databasen. Sistnevnte er en omfattende høykvalitets-informasjonskilde for transkripsjonsfaktorer, tilgjengelig på www.tfcheckpoint.org. Videre presenterer avhandlingen den første konseptuelle demonstrasjon av en komplementær tilnærming til gastrin systembiologi, utviklet ved NTNU, nemlig semantisk systembiologi.

Sett under ett, danner arbeidet som er utført og presentert i denne avhandlingen et solid rammeverk for storskala hypotesegenerering som igjen kan bidra til å oppnå omfattende mekanistisk innsikt i gastrin-indusert genregulering og cellulære responser. Flere av de presenterte tilnærmingene har generell anvendelse for analyser av regulatoriske systemer.

Table of Contents

ABSTRACT

The wish to understand biological complexity has been a great driver for the development of new technologies and knowledge resources. Advanced and quantitative high-throughput technologies have made the generation of enormous amounts of data on a system almost trivial. To acquire knowledge from the data, advanced genome-scale data analysis approaches have become the next bottleneck. The proper care and management of existing (prior) knowledge produced by the global scientific community is pivotal for extracting advanced biological insight from novel high-throughput data.

This doctoral thesis presents the work performed to provide a foundation for systems level analysis of gastrin mediated signal transduction cascades and cellular responses (the 'gastrin signaling system'). It describes the integration of knowledge from a great variety of sources into a network model that captures the knowledge at the time of submitting this thesis concerning the gastrin signaling system, which should impact the way future work can further the comprehensive understanding of gastrin responsive signal transduction and gene regulation networks.

Transcription factors constitute an important link between the signaling cascades and ensuing gene regulatory changes. To provide precise knowledge concerning transcription factors (TFs), this thesis focuses in part on constructing a comprehensive overview of the current knowledge on TFs, both by establishing a robust set of curation guidelines for curating the transcription factors that are experimentally

verified and documented in literature, and by building and making available the TFcheckpoint database: a high-quality and comprehensive information resource on transcription factors, available at www.tfcheckpoint.org. The thesis furthermore presents the first proof of concept of a complementary approach to perform gastrin systems biology developed at NTNU, namely Semantic Systems Biology.

Together, the work carried out and presented in this thesis provides a solid framework for large-scale hypotheses generation to gain comprehensive mechanistic insight concerning gastrin mediated gene regulation and cellular responses. Moreover, many of the approaches are generally applicable for any regulatory system analysis.

ACKNOWLEDGEMENTS

The work in this thesis was conducted in the period of 2009 – 2013 at the Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology (NTNU), Trondheim, Norway. This work was funded by the Norwegian Cancer Society and The Liaison Committee between the Central Norway Regional Health Authority (Samarbeidsorganet).

First and foremost, I am tremendously grateful to my supervisor Professor Astrid Lægreid for excellent guidance during my thesis work. Her extensive knowledge of the subject, enthusiasm, and motivation has been invaluable for my thesis work. Thank you for being extremely supportive and infusing positive energy throughout my thesis work. I also greatly indebted to my co-supervisor Professor Martin Kuiper for his excellent advice, ideas and support throughout my thesis work. I always felt great whenever you organized get together at your place such as 'take-away Indians'. I also want to pay my sincere gratitude to my co-supervisor Professor Liv Thommesen for sharing extensive knowledge, and being extremely supportive and friendly. Thank you for always taking time to answer my questions and quick comments on the manuscripts.

I would like to thank my co-authors and group members; Torunn Bruland, Åsmund Flobak, Konika Chawla, Aravind Venkatesan, Jayavelu Naresh Doni, Vladimir Mironov and Nadi Bar for their excellent work and invaluable feedback. I also thank all my co-authors from abroad; Anaïs Baudot, Karen R. Christie, Rama Balakrishnan, Rachael Huntley, David P. Hill and Judith A. Blake for providing excellent and quick inputs on the manuscripts.

I also thank all my group members for creating pleasurable environment and participating in scientific discussions. I will always remember the friendly banter and tremendous support that my office mates provided.

All past and present colleagues are highly appreciated for creating a pleasant and stimulating working atmosphere.

I thank all my friends and well-wishers. I would like to give special thanks to my close friends Shajrul, Samir, Amit and Anil for their useful advices, persistent support and motivations. My deepest thank goes to my better half Konika Chawla for being extremely caring, supportive and loving.

Finally, I wish to express my gratitude to my family and relatives. I am especially grateful to my parents for being extremely patient and always showing faith in me. Your unconditional love and affection always motivated me as a person. I would also like to thank my siblings and siblings-in-law for their great support and advice.

I would also like to express my sincere thanks to my late grandfather whose vision and motivation was driving force for me to always excel in life.

Trondheim, May 2013

सुशील त्रिपाठी

(Sushil Tripathi)

LIST OF PAPERS

I. **Sushil Tripathi**, Åsmund Flobak, Konika Chawla, Anaïs Baudot, Jayavelu Naresh Doni, Nadav Skjøndal-Bar, Torunn Bruland, Liv Thommesen, Martin Kuiper and Astrid Lægreid. **The Gastrin and Cholecystokinin Receptors mediated signaling network: A scaffold for data analysis and new hypotheses on regulatory mechanisms.** (Submitted to BMC Systems Biology)

II. Konika Chawla*, **Sushil Tripathi***, Liv Thommesen, Astrid Lægreid and Martin Kuiper. **TFcheckpoint: a curated compendium of transcription factors.** (Accepted to Bioinformatics)

III. **Sushil Tripathi**, Karen R. Christie, Rama Balakrishnan, Rachael Huntley, David P. Hill, Liv Thommesen, Judith A. Blake, Martin Kuiper, Astrid Lægreid. **Gene Ontology Annotation of Sequence specific DNA-binding Transcription Factors: Setting the Stage for a Large Scale Curation Effort.** (Accepted to Database)

IV. Aravind Venkatesan*, **Sushil Tripathi***, Vladimir Mironov, Astrid Lægreid and Martin Kuiper. **Network candidate discovery using the Gene eXpression Knowledge Base.** (Manuscript)

*equal contribution

ABBREVIATIONS

AKT	Protein kinase B (PKB)
AP-1	Activator protein-1
ATF	Activating transcription factor
BFO	Basic Formal Ontology
bp	Base pair
CCK1R	Cholecystokinin 1 receptor
CCK2R	Cholecystokinin 2 receptor
CCKR	Cholecystokinin receptor
CCO	Cell cycle ontology
ChIP	Chromatin immunoprecipitation
CHX	Cycloheximide
Clu	Clusterin
CREB1	cAMP responsive element binding protein 1
CRM	cis-regulatory module
CYLD	Cylindromatosis (turban tumor syndrome)
D cell	Somatostatin releasing cell
DAVID	Database for annotation, visualization and integrated discovery
DbTF	DNA-binding transcription factor
EGFR	Epidermal growth factor receptor
EGFR	Epidermal growth factor
EGR1	Early growth response 1
ELK1	ELK1, member of ETS oncogene family
EMSA	Electrophoretic mobility shift assay
ENCODE	Encyclopedia of DNA elements
ERK	Extracellular signal-regulated kinase
FAK	Focal adhesion kinase
FKT	Functional knowledge transfer
FOXO	forkhead transcription factor
G cell	Gastrin releasing cell
GEO	Gene expression omnibus
GeXKB	Gene expression knowledge base
GOC	Gene ontology consortium
GPCR	G-protein coupled receptors
Grb2	Growth factor receptor-bound protein 2
GRN	Gene regulatory network
GSC	Genomic standards consortiums

HPRD	Human Protein Reference Database
HRG	Heregulin
HUPO	Human Proteome Organisation
ICER	Inducible cAMP early repressor
IEA	Inferred from electronic annotation
IGF1R	Insulin-like growth factor 1 receptor
IRS1	Insulin receptor substrate 1
JAK	Janus kinase
KEGG	Kyoto encyclopedia of genes and genomes
MAPK	Mitogen activated protein kinase
Mcl-1	Myeloid cell leukemia sequence 1
MIAME	Minimum information about a microarray experiment
MIAPE	Minimum information about a proteomics experiment
MIBBI	Minimum information for biological and biomedical investigations
MIGS	Minimum information about a genome sequence
MMP	Matrix metalloproteinase
mTOR	Mammalian target of rapamycin
NCA	Network component analysis
NCBO	National Center for Biomedical Ontologies
NFκB	Nuclear factor kappa B
NPC	Nuclear pore complex
ODE	Ordinary differential equation
OWL	Web ontology language
PAI	Plasminogen activator inhibitor
PAK1	p21 protein (Cdc42/Rac)-activated kinase 1
PBM	Protein binding microarray
PDB	Protein data bank
PI3K	Phosphotidylinositol 3-kinase
PKC	Protein kinase C
PLC	Phospholipase C
PPI	Protein-protein interaction
Raf	Murine leukemia viral oncogene homolog
Ras	Rat sarcoma viral oncogene
RDF	Resource Description Framework
RNAP II	RNA polymerase II
RPS6KA4	Ribosomal protein S6 kinase, 90kDa, polypeptide 4
RPS6KA5	Ribosomal protein S6 kinase, 90kDa, polypeptide 5
RSK	Ribosomal S6 kinase
RTK	Receptor tyrosine kinase
RUNX3	Runt-related transcription factor 3

SBML	Systems biology markup language
SELEX	Systematic evolution of ligands by exponential enrichment
SIRT1	Sirtuin 1
Sos	Son of sevenless
Src	Rous sarcoma oncogene
SRF	Serum response factor
STAT	Signal-transducer and activator of transcription protein
TCF7L2	Transcription factor 7-like 2
TF	Transcription factor
TFBS	Transcription factor binding site
TG	Target gene
T-LGL	T-cell large granular lymphocyte
TLR	Toll-like receptor
W3C	World wide web consortium
Y2H	Yeast 2 hybrid

INTRODUCTION

1. Systems biology

Complexity in a biological system arises due to intricate interactions between its different components and the surrounding environment. System-level understanding of these connections can provide detailed view of the biological mechanisms that underlie homeostasis or response to biological signals. "Systems Biology" is concerned with the study of biological function that is derived from interactions [1-4]. 'System' (from the Latin word 'systēma', and this again, is from the Greek σύστημα 'systēma') stands for an entity that maintains its existence through the mutual interaction of its parts [5]. Systems biology is an inter-disciplinary approach which aims at systems-level understanding of biological systems, and as such it involves the representation and analysis of complex biological systems. The roots of systems biology are found in various disciplines such as biology, bio-medicine, biochemistry, computer science/informatics, mathematics, physics, and engineering, and the meaning of systems biology is different for each of these different disciplines. It generates new insight and knowledge which can have qualitative and/or quantitative predictive or explanatory power at the system level [6]. As shown in Figure 1, in this approach iterative cycles of modeling and experimentations are applied which can be used to tackle the challenges faced in understanding the systems level behavior of cells or organisms. According to Mendoza [7], *"systems biology is a transition from qualitative biology towards a quantitative biology, from structural, static descriptions to functional, dynamic properties, and from*

descriptive knowledge to mechanistic knowledge". Systems biology combines reductionist and integrative approaches by which it attempts to decipher and understand the biological meaning of vast genome scale data. The reductionist approach tries to identify and characterize different parts of the system whereas the integrationist approach focuses on investigating interactions between different parts and with their environment. It is an integrated approach that brings together and leverages theoretical, experimental and computational approaches in order to establish connections among central molecules or groups of molecules in order to aid the eventual mechanistic explanation of cellular processes and systems [8].

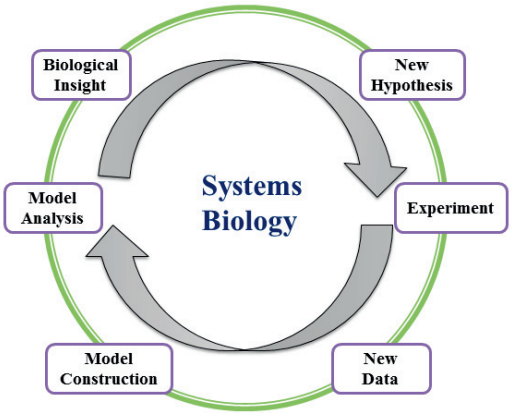


Figure 1: Systems biology cycle. Systems biology allows building, simulating and validating new hypotheses from a systems perspective, taking into account the dynamic interactions of biological parts and processes and the emergence of new functionalities. This is achieved by iterated cycles through a systems biology workflow where genome- or system-wide, or experimental results on systems with interactions between only a few components provide data for the generation of hypotheses which, via adequate hypothesis testing provide an improved understanding of the system as well as improved cycles of experimental work (genome-wide screening or small scale), hypothesis generation and hypothesis testing *Modified from* [9].

In the following sections, I will first give an overview of the model-driven knowledge discovery approaches in systems biology, followed by a description of biological networks that provide biological insight of the intracellular regulatory mechanisms and an introduction to biological databases, integration approaches for biomedical knowledge management. In the last part of the introduction, I provide an overview of the hormone gastrin and its biological significance.

2. Systems biology modeling and approaches for biological discovery

2.1 Modeling in systems biology

Complexity in biological systems mainly arises from large numbers of interactions. A model of these interactions epitomizes knowledge concerning the functionality, structure or behavior of a biological system. There are four basic steps that have to be considered when building a model: construction, verification, calibration and validation [10]. Due to the inability of existing software to perform all these steps, a user has to consider a number of different software packages. Strengths and weaknesses of some of the software packages for model building are discussed by Alves et al. [11]. A computer-based model capturing the systems property can be either qualitative or quantitative in nature.

In a qualitative model, network elements are connected by functional interactions, and each element is characterized by its local state. The states of the components change in time based on their interactions in the network, creating different global states. This type of model mainly represents prior knowledge concerning a biological system in the form of reactants and reactions. Qualitative models are nondeterministic hence they allow many possible outcomes of a chain of events. Today there are several web-based and stand-alone modeling tools for the graphical

representation of qualitative models [12]. CellDesigner [13] is one such tool that has been utilized to create qualitative maps e.g. EGFR mediated signaling [14], mTOR signaling network [15], and AlzPathway [16]. A qualitative model embodies structural properties of the system; and by incorporating temporal behavior in a qualitative model, it is possible to forecast time dependent dynamics of the system. For example, Boolean networks [17, 18] and Petri nets [19]. Each component in a quantitative model is characterized by quantity (e.g. concentration or activity). A quantitative model denotes spatio-temporal characteristic of a system, expressed as quantities or activities over time, and the temporal dynamics of the network as a whole is the essential feature of a quantitative model. A number of different methodologies for quantitative modeling exist [20], such as continuous dynamic models e.g. ordinary differential equation (ODE) approaches, which allows to calculate the change of concentration per unit time for any given species in the model. And, discrete dynamic models that allow dynamic simulations even without exhaustive quantitative information of time dependent variation of the components in a network. Quantitative models can be either deterministic or stochastic. In the latter variant, reactants interact in probabilistic manner. Quantitative models allow simulation of the processes that the model describes as well as have predictive power to generate novel hypotheses.

Experimental validation or falsification of the hypotheses generated by qualitative or quantitative models helps to improve the model. The study by Kholodenko and his coworkers on downstream events of EGFR and MAPK cascade with kinetic modeling and experimental validation approach was a stepping stone in this direction [21-23], which was later adapted and adopted by others [24, 25]. Kholodenko et al. derived a kinetic model by converting each reaction in the EGFR signaling cascade into mathematical equations known as chemical kinetic

equations. This model not only predicts the cellular outcome in response to EGF but also provides detailed insight about conditions that lead to time dependent activation of various signaling components in response to EGFR stimulation [21]. Similarly, Nakakuki et al. [26] constructed an initial ODE model, which was further refined following the experimental validation, to describe the role of c-Fos transcription factor in EGF and HRG mediated cell fate decisions. Zhang et al. [27] formulated a Boolean dynamic model of T cell large granular lymphocyte (T-LGL) survival signaling through which they managed to identify and validate experimentally the potential causes and regulators of the T-LGL survival. Schilling et al. generated a quantitative model that has ability to predict the role of cytokine-receptor mediated ERK activation in cell fate decisions [28]. Similarly, Bianconi et al. constructed a dynamic model of EGFR and IGF1R mediated pathways in non-small cell lung cancer to study the involvement of MAPK cascade in governing migration or proliferation after receptor alteration [29].

2.1.1 Graph based modeling

Functional organization of biological systems is attributed to networks of large numbers of components and their interactions [30]. These networks represent the architectural framework that is associated with cellular decision making [31-33]. Topological analysis of the network can provide very useful information concerning network organization. As illustrated in Figure 2, the basic network features that allow us to study the network topology and characterize complex networks include: a) node degree and degree distribution; reflecting the number of links the node has to other nodes, and the characteristics of all nodes in a system, b) path length; which tells the number of links we need to pass through to travel between two nodes, and c) clustering coefficient, represents the tendency of a network to form clusters of interconnected nodes [30, 34].

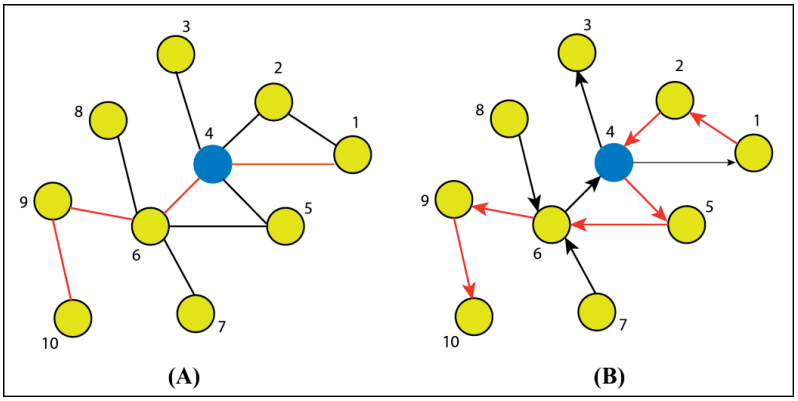


Figure 2: Network topology analysis. (A) An undirected network that includes 10 nodes. The degree k of node number four (blue) in this network is '4'. The 'red' lines represent the network distance between node 1 and 10, which in this case is '4'. **(B)** In the directed version of the network, node number four (blue) is characterized by *in-degree* = 2, and *out-degree* = 3. The network distance (red arrow) to travel between node 1 and 10, in this case is '6'. The clustering coefficient of node four in both the cases is 0.2. *Modified from* [34].

As shown in Figure 3, there are three kinds of graph based network models: i) Random networks; where most nodes are connected by an 'average', or typical number of edges (Figure 3A), ii) Scale-free networks; where networks are characterized by highly non-uniform distribution (following a power-law distribution) which means most nodes have only a few links whereas a few nodes have a large number of edges (Figure 3B). Such highly connected nodes are often called as *hubs*, and iii) Hierarchical networks; where small clusters of densely linked nodes (Figure 3C) are interlinked with other dense clusters by a few hubs, generating a hierarchical network that possesses scale-free property and large average clustering coefficient [30, 34].

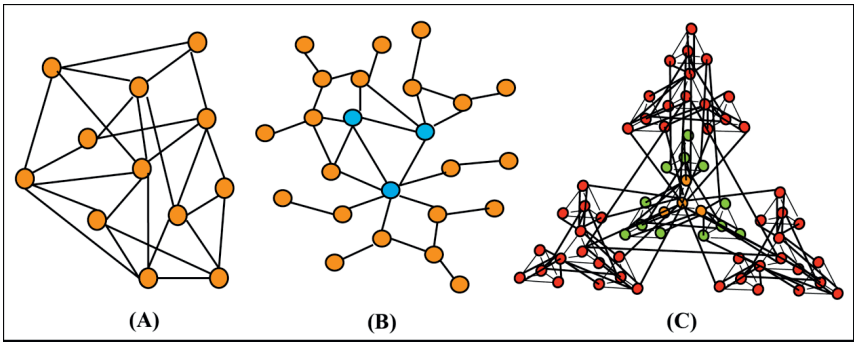


Figure 3: Network models (A) Random network, where node degree follows Poisson distribution indicating that most nodes have approximately same number of links, **(B)** Scale-free networks, characterized by a power-law degree distribution, and **(C)** Hierarchical network, which integrates scale-free topology with large clustering coefficient. *Modified from* [30].

2.2 Biological discovery approaches in systems biology

2.2.1 Bottom–up approach

In bottom-up systems biology, the starting point is structure with low spatial dimensionality such as genes and proteins in an organism. The detailed knowledge about the interactions between these components constitutes the building blocks of bottom-up systems biology approaches (Figure 4). From this knowledge, the modeler aims to reconstruct a network of the system, including feed-forward and feedback regulatory relations [3]. In other words, functionality and behavior of a system are deduced from components of the system that are well characterized with high levels of mechanistic detail.

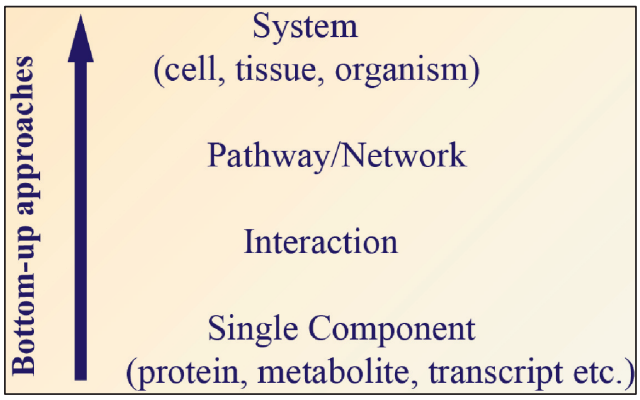


Figure 4: Bottom-up systems biology approach. In this approach, the available knowledge about the components e.g. genes, proteins etc. and their interactions is assembled into a model. This model can be converted into a dynamic model that has capacity to simulate the system behavior in different conditions. This achieved by continuous revision of the model from the knowledge gained through comparison between model-based predictions and their experimental validation. *Modified from* [35].

As illustrated in Figure 5 Bottom-up systems biology studies rely on: *(i) experimental observations that determine the kinetic and physicochemical properties of the components (e.g. enzyme kinetics, diffusion properties) either by studying the components in isolation or by using parameter estimation strategies; (ii) knowledge concerning responses of the subsystem to perturbations while it is in the context of the cell; (iii) the construction of detailed models to calculate the data from (ii) (for model validation and improvement) and to improve experimental design; and (iv) the development of tools for model analysis and representation* [36].

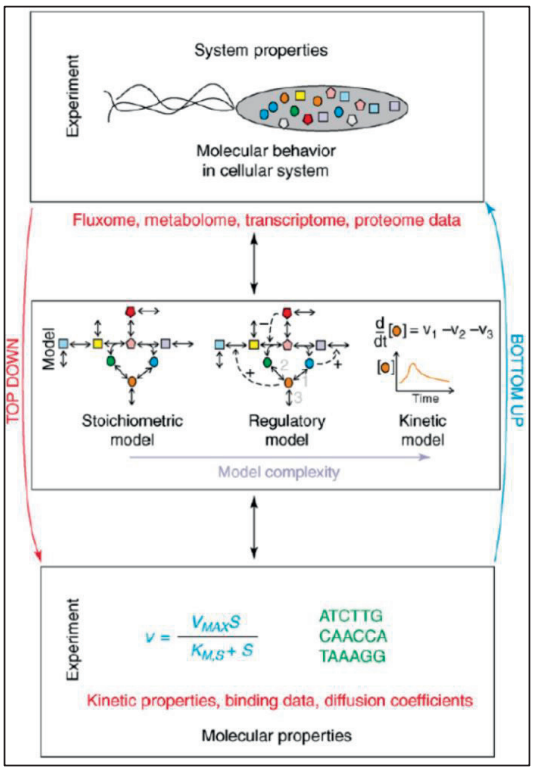


Figure 5: Illustration of Bottom-up and Top-down systems biology approaches. *See text for details. Reprinted from The Lancet, Volume 15, Issue 1, Frank J. Bruggeman , Hans V. Westerhoff, The nature of systems biology, 45 – 50, Copyright (2007), with permission from Elsevier.*

Models are built based on the information available in the literature of specific and sometimes independent experiments. Reconstruction of whole systems with all the mechanistic details from parts of the model or module might work for simple organisms such as prokaryotes but for eukaryotes, organizational complexity at various levels can be a major bottleneck in achieving well-functioning reconstructed whole systems. Often parameter estimation of whole system using small, mechanistically verified module gives false kinetic parameters. It is

more difficult for eukaryotes than for prokaryotes to measure kinetic parameters *in vitro* hence there is a need to develop strategies to measure kinetics of the components *in vivo* [37].

2.2.2 Top-down approach

Top-down systems biology approaches start with experimental data of the system and try to find out its components and their interactions (Figure 6). New hypotheses are generated with the use of correlation and clustering approaches to identify groups of molecules or components which are inter-dependent and/or co-regulated. Comprehensive understanding of a system has led to the generation of large data sets with numerous data points for a single organism, however with limited number of perturbations (e.g. genetic, knock-out, environmental).

A large variety of statistical, pattern finding and classification methods are used to elucidate knowledge from large-scale data [38]. Clustering, which is a type unsupervised learning method that enables class discovery, is a widely used approach [38]. Furthermore, to gain detailed insight from gene expression data, knowledge-based clustering approaches are becoming popular because of the fact that these approaches integrate known information concerning genes from various sources [39]. These approaches are utilized for deducing e.g. groups of co-expressed and co-regulated genes from large gene expression data.

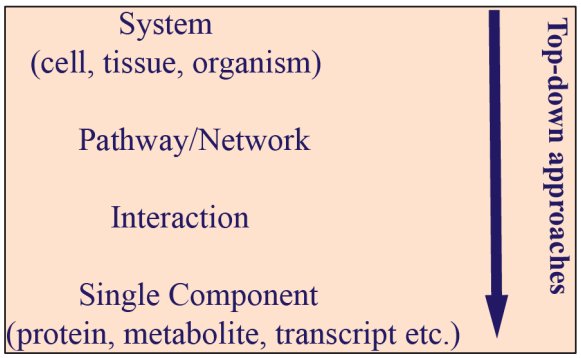


Figure 6: Top-down systems biology approach. In this approach, the starting point is genome-wide data collected from a system that is exposed to different conditions or factors (e.g. growth factors, mutations, nutrients etc.). The data is analyzed using appropriate statistical method to identify components e.g. genes, proteins, or metabolites that exhibit significant change in response to the perturbation. This is followed by clustering and other computational analyses in which structural information e.g. protein-DNA interaction, protein-protein interaction are integrated that can lead to the identification of co-regulated modules. *Modified from* [35].

2.2.3 Middle-out approach

Practical limitations in bottom-up and top-down approaches have led to the adoption of hybrid approaches like the so called "middle-out" approach. Such approaches can be regarded to combine the bottom-up and top-down approaches, as they start somewhere in-between and then work out exploring both 'higher' and 'lower' levels [3, 40]. As represented in Figure 7, systems and functions represent 'higher' levels whereas molecular data and mechanisms depict 'lower' levels. Often, some knowledge about interactions between various players or components of the model is available. Interactions can be partially understood but sufficiently accurate knowledge may still be lacking,

thus significantly reducing the space of the network structure [41]. Thus, from such a middle position (Figure 7) with incomplete knowledge about model components and structure, both higher and lower levels of structural complexity can be uncovered by exploring parameter spaces. This strategy has been followed in the modeling of the heart. For the heart modeling, the starting point was the cellular level modeling of the processes and components that contribute to metabolic or mechanical functions [42, 43]. These cellular models then reached to the tissue and organ level through the incorporation of detailed information of the higher-level structural complexity [44]. In addition, they managed to move from cellular level models to genome level models by modeling the influence of known genetic alterations on the model proteins[45].

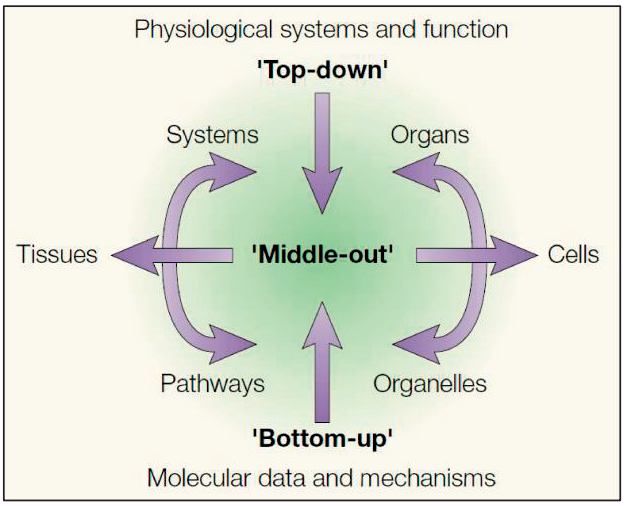


Figure 7: Middle-out systems biology approach. *See text for details. Reprinted by permission from Macmillan Publishers Ltd: Nature Review, Molecular Cell Biology [46], copyright (2002).*

3. Biological networks

Biological networks represent biological insight about intracellular mechanisms gathered through both small-scale and large-scale studies. Among the intracellular networks; gene regulatory, signal transduction and metabolic networks are widely investigated. Since the focus of the present thesis has been towards signal transduction and gene regulation, metabolic networks will not be discussed in detail.

3.1 Protein-protein interaction (PPI) networks

Proteins are the main players of the cellular machinery [47]. In most instances, proteins carry out their biological functions through interaction with several other proteins. For example, the role of androgen and estrogen sex steroids in differentiation is determined by their interacting protein partners, which usually vary with cell types and physiological states [48]. Similarly, the nuclear pore complex (NPC), involved in nucleocytoplasmic shuttling, is estimated to be composed of ~30 distinct proteins (nucleoporins) [49] and engages in a very high number of protein-protein-interactions. In Figure 8, interaction partners of the nucleoporin 210 kDa (NUP210) protein are represented.

Advances in proteomics technologies led to generation of massive amounts of PPI data. PPI based studies have been performed extensively on yeast *Saccharomyces cerevisiae* [50-53] to understand the biological properties resulting from these interaction networks. System-wide PPI detection approaches aim to identify all protein interactions in a system. However, through such approaches there are problems in terms of false positives (i.e. the method reports protein-protein interactions that are not true), as well as missing interactions (false negatives, i.e. true interactions not reported by the method) [54].

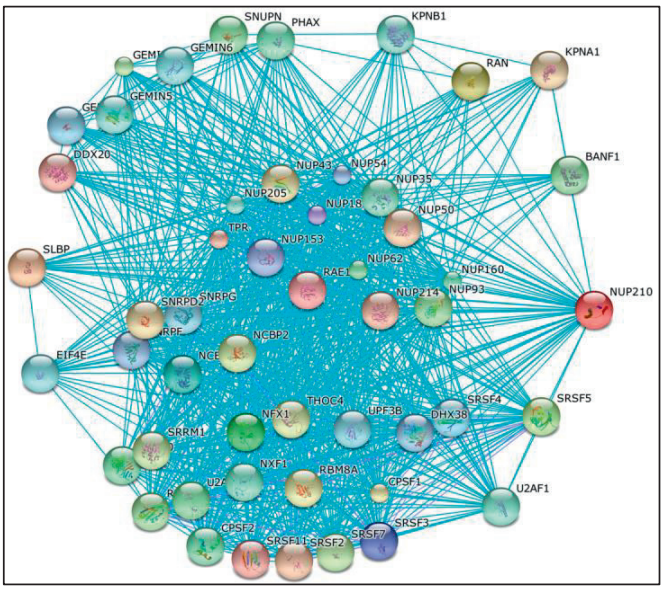


Figure 8: Illustration of nucleoporin 210 (NUP210) protein (red) interaction partners generated using STRING [55] data source.

Currently, several strategies are in practice for large-scale PPI mapping, including:

1) Literature curation of protein interaction data [56]
2) Computational prediction of protein interaction data [57]
3) High-throughput experimental mapping [58]

Each strategy has its own advantages and limitations. Literature curation provides protein interactions derived from small-scale experiments. The observations made from such small-scale experiments are less prone to false positive results. However, literature-curated PPIs have limitation for not being comprehensive. Additionally, due to lack of formalized curation guidelines for recording PPIs from literature there

is variability in the curation process. Similarly, computational prediction methods have the advantage of being applicable at genome and proteome scale in a cost effective manner but are limited by rules for the protein interaction predictions because these rules are not yet precise and exhaustive so that they can reduce the number of false positives and false negatives. From a high-throughput experimental point of view, binary interaction and co-complex interaction methods are widespread in use (Figure 9). Yeast 2 hybrid (Y2H) system based high-throughput method for binary interaction detection, and protein purification followed by identification of constituents by mass spectrometry for co-complex interaction detections are widely used methods for large scale detection of PPIs. Both Y2H and protein purification based methods have their own limitations. As depicted in Figure 9, Y2H may detect interactions which do not occur *in vivo* whereas protein purification based methods may not detect all the interaction partners of a complex.

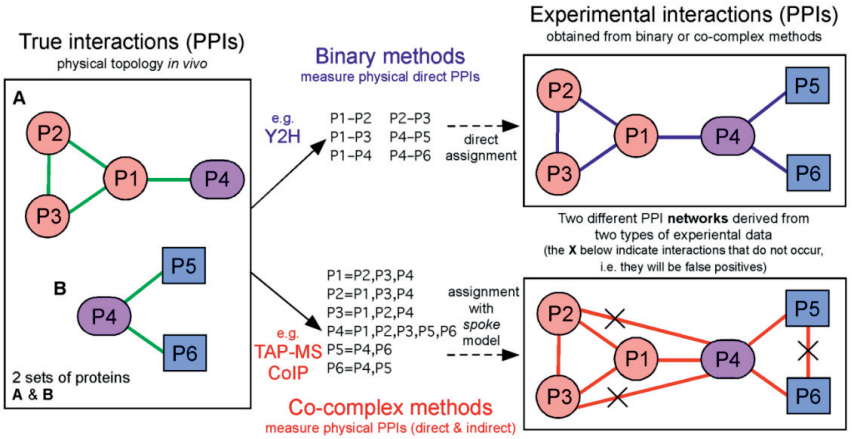


Figure 9: Binary interaction and Co-complex interaction methods to determine PPIs. *Adapted from* [59].

With the continuous increase in PPI data, the repositories that record PPI information (detail list in Table 1) also increased exponentially. Some of the widely accessed PPI databases are the Human Protein Reference Database (HPRD), which is a curated database of human proteins and their interactions [60], IntAct which is an open access resource for molecular interactions derived from literature curation or direct user submissions [61], and iRefWeb that consolidates protein interaction data from 10 different sources [62].

Table 1: PPI databases and resources. *Adapted from* [59].

Acronym	Database Name	URL	PPI sources	Species
Primary Databases: PPI experimental data (curated from published studies)				
BIND	Biomolecular Interaction Network Database	http://bond.unleashedinformatics.com/	Literature-curated	All
BioGRID	Biological General Repository for Interaction Datasets	http://thebiogrid.org/	Literature-curated	All
DIP	Database of Interacting Proteins	http://dip.doe-mbi.ucla.edu/	Literature-curated	All
HPRD	Human Protein Reference Database	www.hprd.org/	Literature-curated	Human
IntAct	IntAct molecular interaction database	http://www.ebi.ac.uk/intact/	Literature-curated	All
MINT	Molecular Interactions database	http://mint.bio.uniroma2.it	Literature-curated	All
MPact	Protein interaction resource on yeast	mips.helmholtz-muenchen.de/genre/proj/mpact/	Derived from CYGD	Yeast
MPPI	Mammalian protein-protein interaction database	http://mips.gsf.de/proj/ppi/	Literature-curated	Mammalian
Meta Databases: PPI experimental data				
APID	Agile Protein Interaction DataAnalyzer	bioinfow.dep.usal.es/apid/		All
iRefWeb	web interface to protein interaction data	http://wodaklab.org/iRefWeb/		All
MPIDB	Microbial protein interaction database	http://www.jcvi.org/mpidb/		Microbial
PINA	Protein Interaction Network Analysis	http://cbg.garvan.unsw.edu.au/pina/		All
Prediction Databases: PPI experimental and predicted data				
MIMI	Michigan molecular interactions	http://mimi.ncibi.org/MimiWeb/main-page.jsp		All
PIPs	Human protein-protein interaction prediction	http://www.compbio.dundee.ac.uk/www-pips/		human
OPHID	Online predicted human interaction database	http://ophid.utoronto.ca		human
STRING	Known and predicted protein-protein interactions	http://string-db.org/		All
UniHI	Unified Human Interactome	http://www.mdc-berlin.de/unihi		human

The protein-protein interaction networks are not only the structural or architectural building blocks but also act as molecular machines. Therefore such networks are one of the active areas of research to understand molecular mechanisms underlying cellular events. This is evident from the fact that the PPIs exhibit emergent properties by performing biological functions beyond the sum of all individual components. For example, the proteasome degradation complex is a collection of ~50 different protein subunits which act together to degrade other proteins [63]. With the technological advances to produce PPI data, network-based applications for the analyses of PPIs also developed exponentially. A list of PPI based analysis and visualization tools are compiled in Table 2.

Table 2: Network visualization tools. *Modified from* [12].

Name	Network visualization	URL
Stand-alone		
Arena3D	PPI	http://www.arena3d.org/
BiNA	PPI	http://www.bnplusplus.org/bina/
BioLayout Express	PPI	http://www.biolayout.org/
BiologicalNetworks	PPI	http://www.biologicalnetworks.org/
BioTapestry	Pathway	http://www.biotapestry.org/
Caleydo	Pathway	http://www.caleydo.org/
CellDesigner	Pathway	http://www.celldesigner.org/
Cytoscape	PPI	http://www.cytoscape.org/
Edinburgh Pathway Editor	Pathway	http://epe.sourceforge.net/SourceForge/EPE.html
GENeVis	PPI	http://www.win.tue.nl/~mwestenb/genevis/
GenMAPP	Pathway	http://www.genmapp.org/
IngenuityPathways	Pathway	http://www.ingenuity.com/
Jdesigner	Pathway	http://jdesigner.sourceforge.net/Site/JDesigner.html
KEGG Atlas	Pathway	http://www.genome.jp/kegg/
Medusa	PPI	http://coot.embl.de/medusa/
MetaCore	Pathway	http://www.genego.com/
NAViGaTOR	PPI	http://ophid.utoronto.ca/navigator/
N-Browse	PPI	http://www.gnetbrowse.org/
Ondex	PPI	http://www.ondex.org/
Osprey	PPI	http://biodata.mshri.on.ca/osprey/
Pajek	PPI	http://pajek.imfm.si/
PathVisio	Pathway	http://www.pathvisio.org/
ProViz	PPI	http://cbi.labri.fr/eng/proviz.htm
SpectralNET	PPI	http://chembank.broad.harvard.edu/resources/
Tulip	PPI	http://tulip.labri.fr/TulipDrupal/
VANTED	PPI	http://vanted.ipk-gatersleben.de/
VitaPad	Pathway	http://bioinformatics.med.yale.edu
Web-based		
ArrayXPath	Pathway	http://www.snubi.org/software/ArrayXPath/
GEPAT	Pathway	http://gepat.sourceforge.net/
Graphle	PPI	http://sonorus.princeton.edu/graphle/
iPath	Pathway	http://pathways.embl.de/
MAGGIE Data Viewer	PPI	http://maggie.systemsbiology.net/
Omics Viewer	Pathway	http://www.biocyc.org/
PathwayExplorer	Pathway	http://genome.tugraz.at/
PATIKA	Pathway	http://www.patika.org/
Payao	Pathway	http://celldesigner.org/payao/
ProMeTra	Pathway	http://prometra.cebitec.uni-bielefeld.de
Reactome SkyPainter	Pathway	http://reactome.org/
STITCH	PPI	http://stitch.embl.de/
VisANT	PPI	http://visant.bu.edu/
WikiPathways	Pathway	http://www.wikipathways.org/

3.2 Signal transduction networks

Only selected stimulatory signals from the external environment are passed through the cell membrane. These stimuli can for instance originate from growth factors, mechanical signals, cell-cell contacts, or nutrients. As shown in Figure 10, signal transduction networks represent a roadmap for the information flow within a cell in response to extrinsic stimuli or intrinsic cues resulting in a number of possible cellular responses such as cell proliferation, migration, and apoptosis. The flow of information through signaling cascades is triggered when stimuli mostly in the form of ligands bind and activate specific cell receptors. Localization of the receptors can be either extracellular or intracellular. Extracellular receptors are integral transmembrane proteins, spanning the plasma membrane, with one part of the receptor outside the cell and the other inside, e.g. G-protein coupled receptors (GPCR), receptor tyrosine kinases (RTKs), and toll-like receptors (TLRs). The ligand e.g. a hormone or a growth factor binds to the outside part of the extracellular receptor and stimulates a cascade of events inside the cell. Intracellular receptors are soluble proteins localized either in the cytosol or nucleus and are also termed cytoplasmic receptors (e.g. NOD like receptors), or nuclear receptors (e.g. steroid receptors, retinoic acid receptors). Most nuclear receptors shuttle between the nucleus and cytoplasm. Ligands which activate these receptors pass through the plasma membrane and bind to the receptor e.g. steroid hormones which bind nuclear receptors.

As depicted in Figure 10, basically there are two stages in signal transduction processes:

a) Extracellular stimulus activates specific receptor protein on the membrane

b) Second messengers and proteins (e.g. kinases, phosphatases etc.) transmit signals inside the cell and elicit cellular responses.

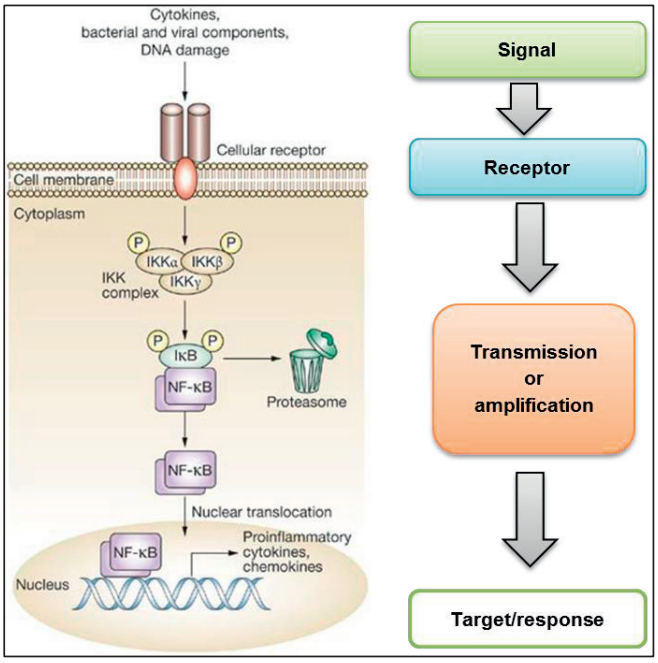


Figure 10: Illustration of signal transduction pathway. Signal transduction cascades are activated in response to extrinsic and intrinsic cues which regulate transcription programs in the cell. Transcription factors are the key components of the transcription regulation that plays a determining role in conferring various cellular responses. However, changes in gene expression are not mandatory for cellular outcomes to occur in response to a signal since signaling components present in the cell at the time point of stimulation can sometimes trigger responses like e.g. secretion from pre-filled vesicles via post-translational mechanisms. *Reprinted by permission from Macmillan Publishers Ltd: Nature Clinical Practice Rheumatology [64], copyright (2007).*

During the course of signal transduction processes usually many proteins and other molecules participate in transducing the information from a given receptor to the cellular responses, thus creating a signal

cascade (Figure 10). Interconnections between signaling pathways results in networks. In a network, a component can receive signals from multiple inputs or signaling pathways. Such components can be regarded as integrators/junctions. Similarly, certain components can be involved in transmitting signals to multiple pathways. These components are considered as nodes that split the signal [65]. Furthermore, components that receive multiple inputs and transmit signals to multiple pathways can act as switches. For enhanced understanding of the intracellular cascades and signaling cross-talks which are linked to various cellular responses, a comprehensive representation of the signaling reactions is crucial. Section 2.1 of this thesis provides a brief account of the literature-curated signal transduction pathways.

3.3 Gene regulatory networks (GRNs)

Gene regulatory networks depict physical and/or regulatory relationships between transcription factors and their target genes. These networks are bipartite and have a functional flow of information from transcription factors (TFs) to target genes (TGs). There are two aspects of the interaction between a TF and TG: one concerning direct binding of a TF to specific sequences in the gene regulatory regions (promoter/enhancer) of a TG and the other pertaining to the regulatory effect of a TF on a given TG (Figure 11a). The regulatory interaction results in positive or negative influence of a TF on transcriptional activity at a given TG. Generally, each TF regulates many TGs and each TG is regulated by more than one TF (Figure 11b).

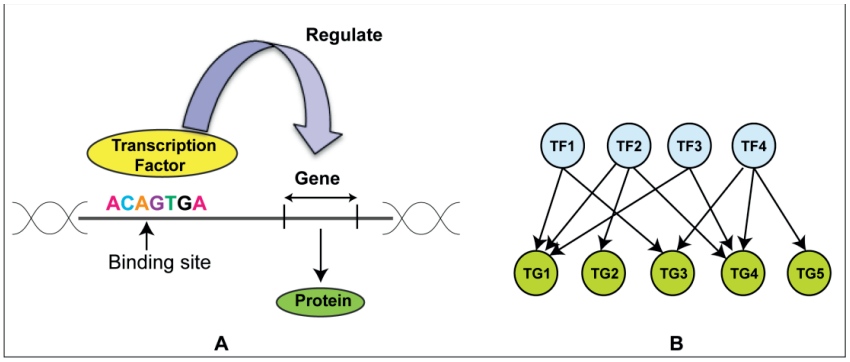


Figure 11: Transcription regulatory network. (A) Sequence-specific DNA binding transcription factors bind to the specific DNA sequence in the regulatory region of the target gene to regulate its expression and further protein systhesis. **(B)** A transcription factor can regulate many TGs, and a TG can be regulated by many TFs.

A variety of methods including both small-scale and large-scale experimental, as well as computational are used to identify specific interactions between a TF and the regulatory regions of its target genes [66, 67]. Some of the experimental methods for the identification of interactions between TF and the TG regulatory region DNA sequences, include EMSA, SELEX, DNA footprinting, protein-binding microarrays (PBM) and ChIP-seq (reviewed in [66]). Computational approaches that are used for identification of interaction of TFs with TG regulatory regions include *de novo* motif finding, as well as algorithms that can identify transcription factor binding sites (TFBS) in genomic regions near genes based on consensus TFBS motifs [66]. MotifLab is one of the freely available online tools for discovering TFBS and cis-regulatory modules (CRM) [68]. CRM is stretch of DNA (100-1000 bp) where number of TFs bind and regulate expression of TGs. Experimental approaches used to identify the regulatory relation between a TF and a

given TG includes reporter gene assay [69], and the experiments performed in whole organism where it is documented that the protein in question binds to the TG regulatory region e.g. measuring the influence of TF on the expression of TG by RNA interference mediated knock-down of TF [70].

Various model organisms ranging from unicellular *E.coli* to the higher non-chordates such as sea urchin have been employed to investigate gene regulatory networks [66]. Each of these model organisms has different advantages and limitations. Davidson et al pioneered the mapping of the complex connections between the gene regulatory components underlying development in the model organism sea urchin [71, 72]. Compared to non-mammalian organisms progress in understanding GRNs with mammalian organism such as mouse and human has been relatively slow. However, the Encyclopedia of DNA Elements (ENCODE) project (http://genome.ucsc.edu/ENCODE/), which is mammalian-centered, was initiated with an objective to decipher systems-level gene regulatory networks across different cell lines and cell types. The ENCODE consortium has produced massive amounts of data relevant for the understanding of DNA regulatory elements, and ongoing analyses of these data is expected to greatly improve an in-depth understanding of the gene regulatory networks [73, 74]. So far, ENCODE has generated 457 ChIP-seq data sets on 119 human TFs in 72 cell lines. They developed a computational method (*de novo* motif discovery) to identify and further characterize sequence-specific motifs of each of these TF in ChIP-seq peaks [75]. Furthermore, analysis of the genome-wide TF binding profiles and TF co-association patterns by Gerstein et al. has provided the beginning of a systems-level understanding of the TF regulatory wiring [76].

4. Knowledge discovery and integration

4.1 Omics data, information and knowledge management

The advent of high throughput experimental technologies revolutionized biological research from a relatively data poor discipline into a data rich. The whole–genome sequencing of *Haemophilus influenza* in year 1995 [77], and the successful completion of human genome sequencing in 2003 [78] revolutionized the omics field. The term 'omics' refers to the totality of a class of data of a biological system under a series of perturbations. Due to profound advancements in high throughput technologies, a variety of omics subdisciplines (genomics, proteomics, metabolomics, interactomics and so on) have begun to emerge [79], generating massive amounts of data for the comprehensive understanding of biological systems and processes. According to Joyce et al. [80]; as illustrated in Figure 12, the molecular information captured by these omics data can be classified into three broad categories:

i) Components data – which yields information regarding the specific molecular content of the cell or systems e.g. genomics, proteomics, transcriptomics and metabolomics.
ii) Interactions data – which specifies the connectivity that exists between the molecular components e.g. protein-DNA and protein-protein interaction data.
iii) Functional states data – which captures overall behavior/phenotype of the biological system e.g. fluxomics and phenomics.

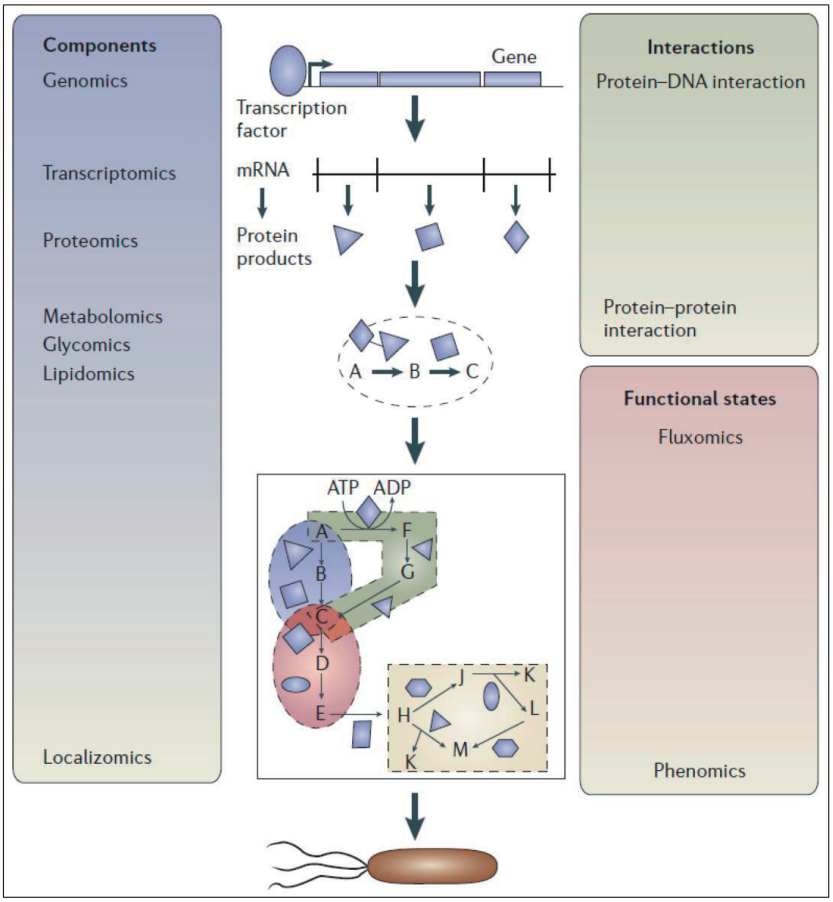


Figure 12: Omics data aim to provide comprehensive information of all components and interactions within the cell. *Reprinted by permission from Macmillan Publishers Ltd: Nature Review, Molecular Cell Biology [80], copyright (2006).*

To gain significant biological insights from omics data, it is mandatory to have structured and standardized data storage formats to enable data sharing between multiple sources, usually enabled through databases where they can be queried, analyzed and compared. The solutions for data reporting standards include minimum information checklists i.e. The Minimum Information for Biological and Biomedical Investigations (MIBBI) [81], controlled vocabularies or ontology terms [82], and standard file formats [83]. The Genomic Standards Consortiums (GSC) published a guideline as 'minimum information about a genome sequence' (MIGS) defining core descriptors for genome and metagenome submissions [84]. Similarly, there are various data standard initiative such as MIAME for microarray [85], HUPO and MIAPE for proteomics data [86, 87], and metabolic standard initiative [88] for metabolomics studies. The next hurdle is to establish efficient and standardized data exchange formats which can aid the integration of data residing in different sources and to provide users an interface to access merged information which is unified and unambiguous. To achieve this, data from different sources must be parsed and merged into unified formats, where the main challenge is to handle different formats from different sources. Today there are many such databases e.g. GeneCards [89-93], which is a compendium of annotated information concerning human gene, whereas BioMart [94] and DAVID [95] are knowledge bases and analysis tools that provide unified view of the data retrieved from one or multiple resources. Similarly, IntegromeDB [96] semantically integrates various molecular biology resources, and provides this integrated information for each searchable gene/protein.

4.1.1 Biological discovery using omics data

The exponential increase in omics data creates major bottleneck for approaches to extract biologically meaningful information from the data. There are several data-driven modeling approaches for signal transduction and gene regulation networks each with its own advantages and disadvantages [97, 98]. Typically, such network modeling methods include Boolean, Bayesian, and neural networks, or models built around differential equations [99]. The network linkages inferred from these approaches may indicate the existence of previously unknown interactions and may thus enable generation of new hypotheses which can subsequently be tested by pertinent experiments.

As depicted in Figure 13, computational approaches to extract biological insight from the high- throughput data generally address three aspects: i) identify the network scaffold by delineating the connections that exists between molecular components of the cell; ii) decompose the network scaffold into its constituent parts, or network modules, in an attempt to understand the overall network structure and iii) develop models to simulate and predict network behavior that gives rise to cellular phenotypes [80].

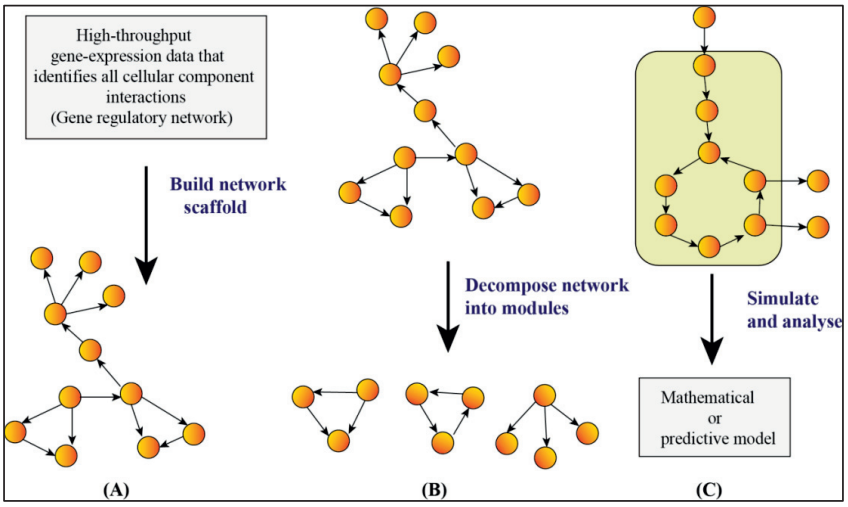


Figure 13: General approaches to investigate the properties of omics data sets. **(A)** Network scaffold identification, **(B)** Network scaffold decomposition, and **(C)** Modeling and analysis of cellular systems. *Modified from* [80].

Networks and pathway visualizations have become routine in the scientific community, to communicate their findings from high throughput data. This is one of the reasons for the continuous increase in the web resources holding networks and pathways. The majority of such resources can be found in a biological pathway portal, Pathguide (http://www.pathguide.org/) [100]. Similarly, in order to visualize and infer biological insights from large data there has been a significant increase in omics data visualization tools too [12].

4.2 Knowledge Bases

Biological knowledge concerns information or knowledge that is documented through scientific research in the life sciences. The existing biological knowledge is the foundation for all biological research. This knowledge is available either in an unstructured form in scientific

publications or in a structured form in biological databases. With the technological advances enabling large scale biological system measurements, there has been a rapid increase in the databases that hold this information. Databases may focus on knowledge about one particular aspect of biology for example UniProtKB [101] and Protein Data Bank (PDB) [102] are protein centric; BioModels [103], Reactome [104], and PANTHER [105] are pathway centric. Similarly, there are different repositories for storing information retrieved through different experimental investigations e.g. the Genomes OnLine Database (GOLD) [106] and GenBank [107] as repositories for genomics related data; GEO [108] and ArrayExpress [109] for transcriptomics; and ENCODE [110] for protein-DNA interactions. With the exponential increase in the number of individual databases it is speculated that by 2015 the number of publications citing 'database' in the title can reach up to 2000 per year [111]. This is evident from the fact that currently there are more than 300 databases that encompass pathway related information [112], about 100 that are concerned with protein structure and protein domain, and over 50 databases that hold transcriptomic information (http://www.oxfordjournals.org/nar/database/c/), and this number is increasing year after year. As a consequence, for each data type there are multiple data resources that present overlapping or only slightly different views of the same data, therefore, a strategy that can provide unified information from multiple resources is required. Systems biology and integrative biology are closely linked. Thus, knowledge management in systems biology addresses the need of knowledge unification through integration of the biological knowledge from various sources that is available in digital formats.

4.2.1 Ontologies and controlled vocabularies

The word 'ontology' originates from metaphysics where ontology represents the branch that concerns with the study of *being* and their relations. According to Gruber [113] "an ontology is an explicit specification of a conceptualization". Ontologies are classified into different types : domain specific ontologies such as Gene Ontology (GO) [114] which focuses on concepts that are relevant for the molecular and cellular biology domain; application ontologies e.g. Cell Cycle Ontology (CCO) [115] which brings together the various concepts concerning cell cycle control; and top-level ontologies e.g. Basic Formal Ontology (BFO) [116] which models common elements that define a generic, integrative framework for essentially all existing domain ontologies (e.g. GO).

Because of the fact that ontologies provide a common vocabulary to support sharing and reuse of knowledge, ontologies provide the foundation for biological knowledge management. To improve communication across different biomedical domains for knowledge management, a repository of biomedical ontologies or controlled vocabularies was created under the auspices of the Open Biomedical Ontologies (OBO) [82]. In addition, ontology-based web services of the National Center for Biomedical Ontologies (NCBO) became available (http://bioportal.bioontology.org/), which facilitated the biomedical community to automatically annotate knowledge with biomedical ontologies. For this, NCBO has developed BioPortal [117], a web portal that enables access to biomedical ontologies developed in variety of knowledge representation formats such as OBO, Web Ontology Language (OWL) [118], and Resource Description Framework (RDF) [119]. The NCBO web services can be conveniently incorporated into software applications to access ontology contents. This feature has

allowed the development of numerous data annotation applications that use NCBO web services, e.g. ISAcreator [120] which supports automatic annotation of experimental metadata.

4.2.2 Data annotation

In the present information rich era, one of the major challenges for the knowledge sources is to provide high-quality and up-to-date information. Annotation of the biological information contained in knowledge sources such as Gene Ontology [114], UniprotKB [101], IntAct [61], KEGG [121], Reactome [104], Entrez gene [122], Ensembl [123] and many more is therefore of high importance. An annotation reflects a connection between an entity and an ontology term assigned to that entity. This connection is created on the basis of inferences drawn from the interpretations that a curator can make from a scientific publication [124]. Experimental analyses or a variety of computational analyses (structural, sequence-based, and other) are some of the methods that can support the curation. Knowledge sources that provide annotated information ideally specify their own guidelines for data annotations. For example, the Gene Ontology Consortium (GOC) provides guidelines for creating annotations to gene products with specific GO terms based on observations and inferences drawn from the experiments, author's statement, or structure and sequence similarity based computational analysis. For each annotation, GO evidence codes describe the type of observation method (e.g. type of experiment (direct assay, mutant phenotype), author statement, computational analysis) that was used for a GO annotation.

Curation is the process of creating annotations of data from the scientific publications.

There are two methods of creating annotations:

1) *Manual curation, where knowledge bases employ human curators who read scientific publications and create annotations*

from that publication. Manually curated databases are generally regarded to contain high quality information but sometimes questions have been raised regarding their completeness as well as on quality [125].

2) *Computationally assigned annotation approaches where computer programs are designed to generate data annotations*. These programs are created around certain rules which are tested and validated against published data or knowledge. The rules can be created to generate annotations not only from the scientific publications but also by searching sequence or structure similarity. The UniProt consortium has developed an automatic annotation pipeline that uses InterPro to automatically annotate UniProtKB/TrEMBL protein entries. Similarly, GOC uses the Inferred from Electronic Annotation (IEA) evidence code for annotations that are transferred automatically either from a database or based on sequence similarity matches that are not reviewed by a curator. Text mining is an emerging field in the direction of creating computer assigned annotations. In this approach computer algorithms automatically extract information from the scientific literatures. This approach has been implicated for extracting information concerning pathway modeling using PathText [126, 127], and for mining the transcription regulatory events [128].

4.3 Data integration

There is no single approach for data integration. Technological solutions to production of a large variety of biological data have led to an increase in the databases storing this information. Thus, to gain a comprehensive view on their data, users must go through a large number of different existing databases harboring relevant information for a given data type. Therefore, it is of utmost significance to unify

information that presently resides in different resources at one place through data integration. Integrated data sources will assist researchers in biological discovery by providing larger and wider overview of the data. Efforts such as BioMart [94], KEGG [121], DAVID [95] and KA-SB [129] are already in place that allow a user to access integrated information. Some popular data integration approaches are described here [130]. Data integration faces some key challenges by virtue of the complexity in various kinds of data resources. For successful integration of the data resources some of these challenges such as common identifier, name, reporting standards, shared semantics, and data curation need to be addressed [130]. As a solution for some of the data integration challenges the World Wide Web Consortium (W3C) founder Tim Berners-Lee coined the term 'Semantic web'. According to the W3C (http://www.w3.org/2001/sw/), "*The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries.*" BioGateway [131] is one of many semantically integrated knowledge resources, which is built on an RDF store that allows access of biomedical ontologies and biological resources. Thus, BioGateway aggregates OBO foundry [82], CCO [115], NCBI taxonomy [132], Swiss-Prot [101], and GO annotation [133] data resources. BioGateway serves as a single source for querying semantically integrated information from these resources through SPARQL [134]. Similarly, there is Gaggle [135] that provides a framework for data exploration and analysis between different software tools and databases within the systems biology community, and Galaxy which is a web-based platform that provides for the integration of several data sources e.g. BioMart [94] and InterMine [136] together with the data analysis tools, has its own data type converters that handles tool specific data format conversion.

5. Gastrin Biology

5.1 Gastrin hormone overview

Gastrin is a gastrointestinal peptide hormone primarily synthesized and released by G cells of gastric antrum (Figure 14). In humans, gastrin is encoded by a gene located on chromosome 17q21. Biologically active gastrin processed via multiple steps with the help of proteolytic enzymes has a C-terminal pentapeptide amide and sulfation of the tyrosine at position 7.

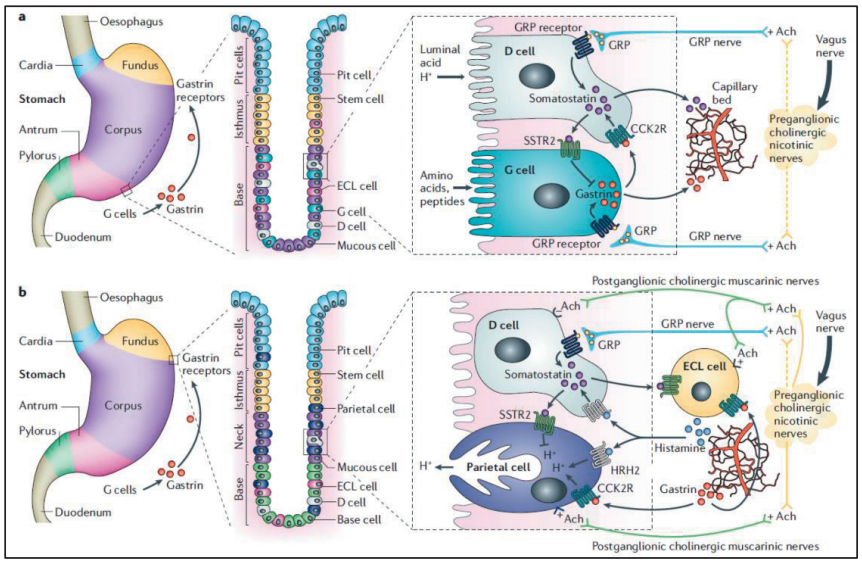


Figure 14: Illustration of cellular interactions of the human gastric mucosa. (a) Cellular interactions of the corpic/fundic mucosa, and **(b)** the antro-pyloric mucosa. *Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Cancer[137], copyright (2006).*

Gastrin exerts its biological functions by binding to the G protein-coupled receptor, CCK2R. CCK2 receptors are located on multiple cell types in the central nervous system, and peripheral organs including stomach (reviewed in [138]), pancreas, and gall bladder [139]. CCK2R-

positive cells in the stomach are enterochromaffin-like (ECL) cells and parietal cells in the corpic/fundic mucosa; (Figure 14b) and D-cells which are in both antro-pyloric and corpus/fundic part of gastro-intestine (Figure 14), Gastrin concentration fluctuates in response to a meal. The pH of the stomach influences gastrin release. High gastric acid (low pH) in the stomach inhibits gastrin secretion from G cells of the gastric antrum (Figure 14a) and low gastric acid (high pH) promotes gastrin secretion. As shown in Figure 14b, gastric acid secretion by parietal cells is mainly controlled by histamine, which is synthesized and secreted by ECL cells of the corpic/fundic part of the stomach and thus functions in a paracrine manner. Figure 14b shows how the activity of the ECL cells to secrete histamine is controlled by gastrin (acting in an endocrine manner since it is produced in a different part of the stomach mucosa) and somatostatin (acting in a paracrine manner, since it is produced by D-cells in the corpic/fundic region) (Figure 14b). Gastrin and somatostatin show antagonistic effect on ECL cells. On the one side, gastrin act as a stimulator, and on the other side, somatostatin sends inhibitory signals to the ECL cells.

The hormone gastrin plays a role in regulation of growth and differentiation of gastric and colonic mucosa [140]. The scientific interest in this hormone is however strengthened by its implication with several diseases. A study on transgenic mice overexpressing human gastrin concluded that hypergastrinemia promotes gastric atrophy, and progression towards gastric cancer [141]. However, in contrast to the above findings, Zavros et al. demonstrated that chronic gastritis in gastrin-deficient mice progresses to gastric adenocarcinoma [142]. It is believed that gastrin overexpression leads to a boost of mitogenic effects, and that it thus promotes the progress of cancer development, whereas, in case of gastrin loss, bacterial overgrowth in the stomach due to hypo-acidity can be a driver of cancer development [137]. In line

with the proposed role of increased levels of gastrin in carcinogenesis, infection with *Helicobacter pylori* has been shown to increase the expression of gastrin [143], and gastrin is considered to be one of the risk factors in *H. pylori* driven gastric carcinogenesis [144].

5.2 Gastrin mediated cellular responses

Gastrin is suggested to affect several cellular responses including proliferation, migration and apoptosis. Like many other extracellular signals such as hormones, growth factors and neurotransmitters acting via GPCRs, gastrin leads to activation of multiple signaling pathways and transcription factors for the regulation of target genes. Many of the cellular responses are mediated through the expression of target genes (reviewed in [138, 145]). Cell line model systems (see Table 3) have been the main tool for exploring the underlying molecular mechanisms of these responses. Nevertheless, some of these finding are also verified in experimental animal model systems such as transgenic mice overexpressing gastrin (INS-GAS) [146] or transgenic mice expressing human CCK2R (ElasCCK2R) [147]. In the following section, I will provide a brief account of the regulatory mechanisms that are pivotal in the regulation of gastrin mediated cellular responses.

Proliferation

Gastrin is the most important trophic hormone of the stomach. One of the main physiological functions of gastrin is to regulate gastric mucosal growth and intestinal epithelial cell proliferation (reviewed in [145, 148]). Through whole animal and cell line studies, it is well established that gastrin stimulates ECL cell proliferation in the stomach [149-151]. Similarly, growth promoting effect of gastrin is also reported in many gastro-intestinal cancer cell lines including rat pancreatic adenocarcinoma cells, AR42J [152], and human stomach cancer cells, AGS [153]. Figure 15 illustrates that gastrin after binding

to its receptor CCK2R, regulates proliferation through activation of multiple signaling cascades such as MAPKs [154-157] and small GTPases [158, 159] , and through regulation of effector proteins such as cell-cycle regulator, cyclin D1 [160]. In AGS cells, gastrin dependent induction of cyclin D1 gene expression is mediated via ß catenin/Transcription factor 7-like 2 (TCF7L2) and CREB [160]. Inducible cAMP early repressor (ICER) is regarded to be a negative feedback inhibitor of cyclin D1 expression and this has been observed in AR42J cells in response to gastrin [161] (Figure 15). In contrast to the growth promoting function of the gastrin in AR42J and AGS cells [152, 153], Muerkoster et al. found that gastrin suppresses growth by inducing apoptosis in colon cancer cells (Colo320) transfected with CCK2 receptor [162].

Apoptosis

Apoptosis plays an important role in cell homeostasis in the gastrointestinal tract [163]. Gastrin is reported to exert both pro- and anti-apoptotic functions [162, 164-166]. A study on hypergastrinemic mice concluded that gastrin induced apoptosis of gastric cells contributes to gastric carcinogenesis [167]. Another finding suggests that hypergastrinemia increases susceptibility of gastric cells to undergo apoptosis [164]. As shown in Figure 15, gastrin is documented to promote anti-apoptosis through inactivation of pro-apoptosis mediator proteins such as forkhead transcription factors (FOXO1 and 3) and BCL-family proteins [166, 168]. It has been shown that gastrin induces anti-apoptosis through MAPKs, NFκB, PI3K, AKT1, and Rho GTPases dependent regulation of cell survival proteins such as Mcl-1, survivin and clusterin [159, 166, 169-171].

Table 3: List of cell line model systems that have been utilized for investigating gastrin mediated responses.

cell type	species	source
GH3	rat	pituitary
AR42J*	rat	pancreas
IEC-6	rat	small intestine
IMGE-5	rat	gastric epithelial
RGM1	rat	gastric mucosa
colonic epithelial cells	rat	colon
Rat PSC	rat	pancreas
Rat-1	rat	fibroblast
RIE-1	rat	small intestine
RGaR9	rat	gastric mucosa
AtT-20	mouse	neuroendocrine tumor
MC-26 CRC cells	mouse	colon carcinoma
NIH-3T3	mouse	embryonic fibroblast
COS-7*	monkey	kidney
AGS*	human	adenocarcinoma of the stomach
colo320*	human	colorectal carcinoma
DLD-1	human	colorectal adenocarcinoma
HEK293	human	embryonic kidney
HUVEC	human	umbilical vein
KATO-III	human	stomach carcinoma
MKGR26	human	gastric cancer
OE33	human	oesophagus
Panc1	human	pancreas
Caco-2	human	colorectal adenocarcinoma
HCT-116	human	colon carcinoma
HT-29	human	colon carcinoma
SW-480	human	colon carcinoma
SIIA	human	gastric cancer
CHO*	hamster	ovary
InR1G9	hamster	pancreatic cancer
MDCK	dog	kidney

**frequently used cell-lines for investigating gastrin mediated responses. This is based on their documentation in at least five gastrin related scientific publications.*

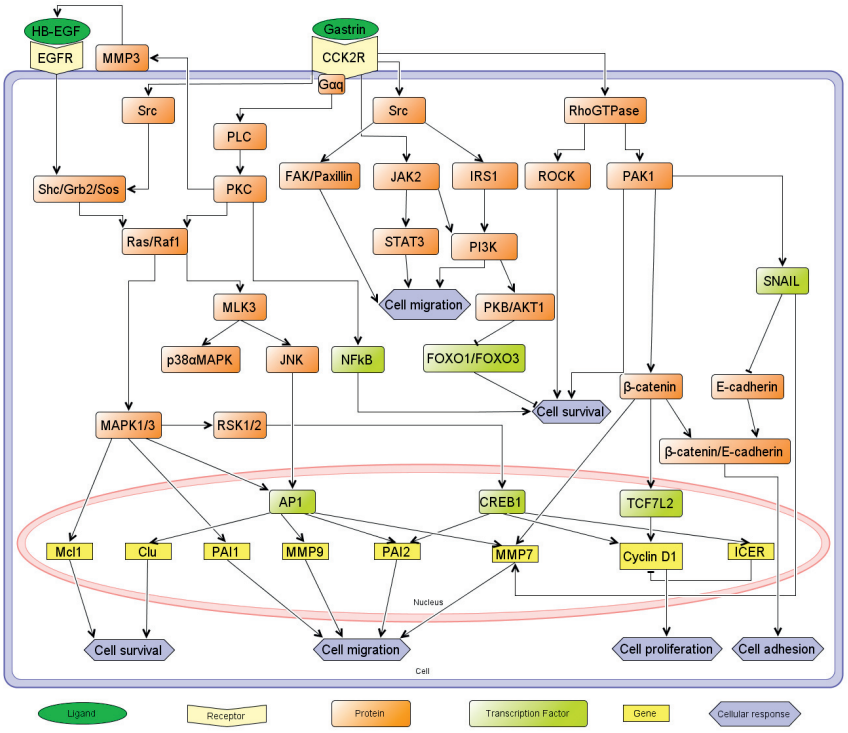


Figure 15 Gastrin mediated cellular response signaling pathways. Gastrin binds to CCK2R receptors and activates multiple signaling cascades. MAPKs are central in regulating gastrin activated cell proliferation, migration, and survival responses. Gastrin dependent activation of MAPKs can follow *PLC/PKC/Ras-Raf* and/or *PKC/MMP3/EGFR/Grb2-Sos* route. *JAK2/STAT3, Src-FAK/Paxillin* and *IRS1/PI3K*, and *PAK1/E-cadherin-β catenin* signaling are involved in the regulation of cell migration while *PI3K/AKT1* and *Rho GTPase* signaling are central in regulating gastrin dependent cell survival (anti-apoptosis). *See text for more details.*

Migration and invasion

Gastrin has been reported to promote disruption of adherens junctions formed upon the interaction of ß-catenin with E-cadherin and α-catenin in intestinal epithelial cell culture experiments [172] (Figure 15). Loss of cell adhesion to the extracellular matrix increases cell motility and invasion which are thought to be linked to promotion of carcinogenesis. As evident from Figure 15, multiple signaling pathways are reported to be associated with gastrin mediated migration and invasion for example MAPKs [173, 174], JAK2/STAT3 [172], FAK/Paxillin [175, 176], and Rho GTPases [159]. In human stomach cancer cells, gastrin is reported to promote migratory responses by augmenting the expression of matrix metalloproteinases, MMP7 and MMP9 [173, 174, 177]. MAPKs, AP1, and ß-catenin signaling pathways are central in gastrin dependent regulation of MMPs [173, 178]. Similarly, gastrin mediated activation of plasminogen activator inhibitor-1 (PAI-1) and plasminogen activator inhibitor-2 (PAI-2) proteins via multiple signaling cascades are documented to be involved in promoting AGS cells migration [179, 180].

OBJECTIVES OF THE STUDY

In modern molecular medicine, a major challenge concerns development of new strategies for knowledge management, including new methods for generation of predictive models that capture essential laws, patterns and principles of biological systems and incorporate experimental data. Biological systems are extremely complex, representing significant modeling and simulation challenges. Model-based systems understanding is also a prerequisite for the development of improved diagnosis and prognosis as well as for identification of new drug targets for e.g. tumor treatment.

The principal objective of the work presented here is to contribute to a comprehensive understanding of the cellular processes involved in carcinogenesis in the gastrointestinal tract, with a particular focus on gastrin-mediated responses. The component objectives are to:

1. Contribute to model-based reasoning for gastrin responses in the form of a complete computer-readable map of the gastrin response signaling pathways augmented by integration of protein-protein interaction data.
2. Provide high quality information resources for experimentally documented mammalian transcription factors by establishing adequate database resources as well as detailed and user-friendly guidelines for Gene Ontology curation.
3. Contribute to better management of knowledge pertaining to gene expression processes in the form of a knowledge base for enhanced reasoning on gene regulation networks, and demonstrate how these knowledge bases can help in knowledge discovery.

SUMMARY OF THE PAPERS

Paper I

The Gastrin and Cholecystokinin Receptors mediated signaling network: A scaffold for data analysis and new hypotheses on regulatory mechanisms

The aim of this study was to construct a literature-curated map of cholecystokinin receptors 1 and 2 (CCKR) signaling pathways mediating biological responses to gastrin and cholecystokinin. We extracted information from more than 250 scientific publications and used CellDesigner software (http://www.celldesigner.org/) to build a comprehensive map that encompasses 519 molecular species including 214 proteins, which are connected by 424 reactions. The map reflects biological understanding extracted from scientific publications and pathway databases. The comprehensive map was curated using the community curation platform Payao. Next, we performed network topology analysis of the CCKR map and identified potential central regulators of the CCKR signaling cascades which include AKT1, SRC, PKC, PAK1, GTPase and HRAS. Furthermore, with the help of BiNoM, we decomposed the CCKR map into sub-networks that can be represented in 18 modules. Each module represents higher level structures and provides enhanced understanding of intracellular processes involved in cellular decisions. In addition, we predict new candidate regulators of CCKR signaling by integrating comprehensive map with large scale protein-protein interaction data. This integration provided us with more than 4000 proteins as novel interactors of the comprehensive network components, including a subset of ~100 interactors that significantly increase the connectivity of the signal transduction network, indicating their potential roles as new regulators of gastrin and cholecystokinin signaling. Using Network Component

Analysis (NCA) informed by gastrin high-throughput time series gene expression data, we generated transcription factor activity profiles that illustrate the dynamic behaviour of the CCKR-regulated transcription factor networks. In this work, we demonstrate how a computational model of complex biological processes such as signal transduction and gene regulation can be integrated with multiple dimensions of large scale data acquisition and analysis thus represents a source for new hypotheses and experimentation to further improve our understanding of CCKR-mediated processes.

Paper II

TFcheckpoint: a curated compendium of transcription factors

The objective of this work was to establish a repository for mammalian RNA polymerase II (RNAP II) regulating sequence-specific DNA binding TFs (DbTFs) that are documented in the scientific literature. A DbTF by definition binds to specific DNA sequences and regulates the RNAP II mediated transcription of the gene that it binds to. We compiled a list of 3462 proteins from 9 major transcription factor database sources for the purpose of curating them with literature evidence. The Gene Ontology term *sequence-specific DNA-binding RNA polymerase II transcription factor activity (GO:0000981)* was selected as the minimum defining term for qualifying a protein as an RNAPII regulating DbTF. We checked for specific scientific publications that would contain evidence to qualify TFs according to our DbTF annotation criteria. Using these criteria we found literature evidence supporting DbTFs for 983 proteins of a total of 3462. Results of our annotation efforts were made available through the TFcheckpoint database (www.tfcheckpoint.org). Entries in this database can be queried by Entrez ID, UniProt ID, gene symbol, and gene name. TFcheckpoint

provides a useful resource for the large-scale gene regulatory network based analyses.

Paper III

Gene Ontology Annotation of Sequence specific DNA-binding Transcription Factors: Setting the Stage for a Large Scale Curation Effort

The purpose of this work was to extend on our work in paper II and establish rigorous guidelines enabling user-community driven creation of Gene Ontology annotations for mammalian sequence-specific DNA binding transcription factors (DbTFs) based on experimental evidence in scientific publications. We devised a framework for using the controlled vocabularies defined by the Gene Ontology Consortium to curate DbTFs based on experimental evidence reported in literature, and to provide an overview of Gene Ontology terms for DNA binding, transcription regulation and transcription factor activity that are eligible for creating DbTF specific annotations. In addition, we describe how to use TF-binding and TF-binding TF activity terms to capture the activity of a TF that is dependent on an interaction with another TF. To contribute to a uniform and well-structured documentation of the experimental evidence for DbTF Gene Ontology annotations, we compiled a list of experimental assays documenting DNA binding, transcription regulation and TF-binding, discussed the eligibility of each of these assays for GO annotations, and indicated how the assays translate into GOC experimental evidence codes. In this work we also describe the strategy to record information reported in literature regarding the target gene that is regulated by the transcription factor that is being annotated. The annotations created using these curation guidelines are made available to the Gene Ontology Consortium and are also stored in our publicly available TFcheckpoint database, together with detailed experimental assay information. Today, GOC

holds experimental evidence-based annotations for 202 mammalian DbTFs. With the initiative describe here, we expect to enrich the GO database with an additional ~600 DbTF annotations that are verified experimentally at the time of the thesis submission.

Paper IV

Network candidate discovery using the Gene eXpression Knowledge Base

In this paper our main goal was to establish a number of use cases of a knowledge base that provides integrated knowledge concerning gene expression events. For this, we established and used The Gene eXpression Knowledge Base (GeXKB), which was built on three application ontologies that capture knowledge concerning gene expression. The GeXKB semantically integrates data from GOA, the IntAct database, KEGG, PAZAR, UniProtKB and NCBI Gene. The GeXKB-contained knowledge was utilized to formulate hypotheses on gastrin mediated regulation of CREB1, TCF7L2, and NFκB transcription factors. We evaluated the hypotheses obtained from GeXKB for further experimental validation based on 1) their documentation in scientific publications as well as on 2) their relevance for gastrin mediated gene regulation assessed from our in-house gastrin time series transcriptomic observations in the AR42J cell line model system. This work demonstrated the value of semantic knowledge bases for knowledge discovery.

DISCUSSION

How a cell responds to external signals (e.g. hormones, growth factors) depends on the state of the cell as well as on information received through the signaling cascades. The information transmission process through these cascades often stimulates several distinct but interrelated cellular responses. These cellular responses are associated with changes in the gene-expression that is controlled by transcriptional regulatory networks. Thus, to unravel the mechanisms governing gastrin mediated cellular responses, we performed extensive genome-wide microarray 14h time-series experiments on gastrin treated AR42J cells. These analyses provided us a list of ~2000 gastrin responsive genes [181]. High-throughput genomic data provides information concerning cellular processes and the intricate connections between different components that are responsible for controlling these processes. Similarly, our gastrin treated microarray time-series data have the potential to depict information about the components and their interactions that are involved in regulating the gene expression changes in response to gastrin during the 14h time period covered in our study. Detailed prior knowledge about the components and their roles in a system can be a powerful tool for deciphering the information encoded in large scale data [182]. The potential significance of the background knowledge led to the development of tools that automatically extract information from structured databases e.g. DAVID [95]. We used this tool in our analyses of the gastrin responsive genome-wide transcriptome data. Another type of initiative aimed at developing tools to access background knowledge is BioCreative (http://www.biocreative.org/), which is a community-wide effort for enhancing the background information extraction process from

literature through text mining [183]. As illustrated in Figure 16, networks reconstructed on existing knowledge provide important scaffold for the expansion of knowledge [184]. However, we found that none of the currently existing resources provide such a comprehensive map of the gastrin response signaling pathways. To elucidate the regulatory mechanisms underlying differential regulation of the gastrin responsive genes it is of immense significance to have a detailed map of gastrin response signaling pathways depicting the flow of information from the gastrin receptor to cellular responses.

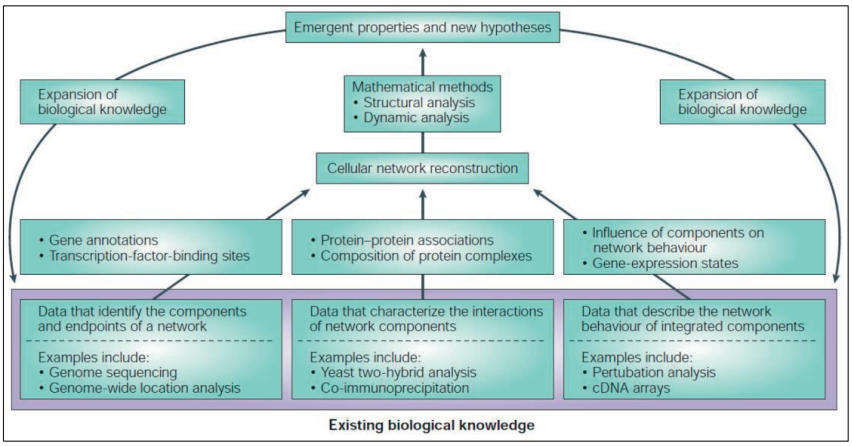


Figure 16: Expansion of biological knowledge through network reconstruction from existing biological knowledge. *Reprinted by permission from Macmillan Publishers Ltd: Nature Review, Molecular Cell Biology [184], copyright (2005).*

Currently, detailed signaling cascade maps providing comprehensive understanding of intracellular networks are available for e.g. EGF receptor [14], toll like receptor [185], gonadotropin releasing hormone receptor [186], and follicle-stimulating hormone induced signaling [187]. We have taken interest in these resources since we regard them to be

valuable examples for addressing the types of challenges we work with. Therefore, we made an effort to construct a similar literature curated comprehensive network of the signaling cascades that mediate cellular responses to Gastrin- and Cholecystokinin-receptors CCK1R and CCK2R (Paper I). The map reflects the biological background knowledge collected from more than 250 scientific publications (until September 2012) concerning CCK1R and CCK2R signaling and is constituted of 519 molecular components that are connected by 424 reactions. This map provides comprehensive signaling knowledge in both computer readable SBML format as well as a high quality graphical representation format that enables a visual comprehension of detailed as well as higher level structures of the network. Using Payao [188], which is a community-based curation platform, we assured high quality network representation by enabling detailed scrutiny of the CCKR network by curators within our research group. This joint effort resulted in a number of new knowledge entries, which were subsequently implemented in an improved representation of the network. We then published the CCKR map in Payao as open source for the scientific community. Thereby we hope to receive comments and tags from the world-wide community of curators in order to use these to keep increasing the quality of the CCKR signaling pathway map and keep it up to date with our increasing biological understanding. A workflow of the CCKR pathway construction and community curation is shown in Figure 17.

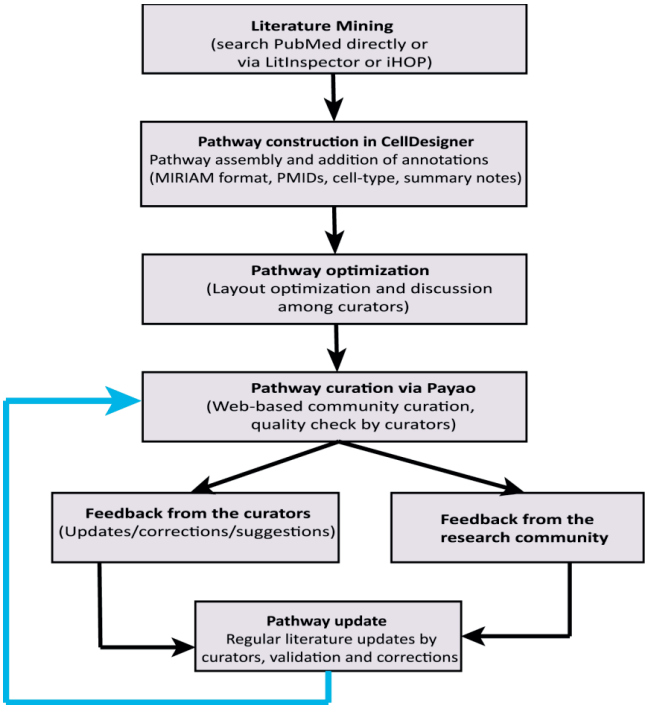


Figure 17: Comprehensive CCKR pathway construction workflow. This workflow summarizes the main steps of CCKR pathway reconstruction from literature, and the quality check using the web-based community curation platform Payao. *Modified from* [189].

In order to improve both the visibility and usability of the CCKR map for the wider scientific community, our CCKR map will be integrated in the Reactome knowledge base. Reactome [104] is a high-quality, curated pathway information resource made available by the European Bioinfomatics Institute, and we are collaborating with Reactome curators to submit the CCKR network.

Signal transduction and gene regulatory networks are central in controlling gene expression changes in response to a stimulus. Gastrin-

mediated intracellular signal transduction pathways influence the expression of a high number of genes by regulating the activity of the transcription regulators in the gene regulatory network. The transcription regulators can be sequence-specific DNA binding transcription factors (DbTFs), and co-TFs (e.g. co-activators, co-repressors, and chromatin remodeling proteins like histone modifiers) [190, 191]. DbTFs and factors that lack DNA binding (co-TFs) but exert their regulatory influence by interacting with DbTFs are central determinants of diverse gene-expression behavior. Among ~2000 gastrin responsive genes, virtually any gene that is classified as a transcription factor can be a key regulator in secondary gene responses underlying gastrin mediated cellular outcomes. Secondary responses are characterized by their dependence on *de novo* protein synthesis (i.e. new transcription-translation of transcription factors regulating them) whereas primary responses are independent of *de novo* protein synthesis (i.e. responses that proceed through post-translational modifications). In order to identify which of the gastrin-responsive genes that acts as transcription factors, we searched for resources that can provide comprehensive and high quality information about transcription factors. Today, there are several mammalian transcription factor information resources e.g. TFCat [192], AnimalTFDB [193], and TFe [194], however, as with many other domain-specific knowledge sources, they either lack sufficient documentation regarding their source of evidence, or lack completeness. To address this challenge, we compiled a list of 3462 candidates for transcription factors (including co-TFs) from major TF knowledge resources and embarked on an effort to record mammalian sequence-specific DNA binding transcription factors (DbTFs) whose functionality is documented in scientific publications either experimentally, or by sequence/structure similarity analysis, or as author's statement. The 983 mammalian DbTFs resulting from this

literature survey are made available through the TFcheckpoint database (http://www.tfcheckpoint.org/) (Paper II). TFCat [192], which is the most comprehensive of the other publically available literature curated transcription factor information resource, contains only 892 of these 983 DbTFs. Thus, to our knowledge TFcheckpoint is currently the most comprehensive and best referenced mammalian TF database. Due to its high quality and comprehensive information it can be a valuable resource for mammalian transcription factor centric studies.

In order to render the DbTF-information contained in the TFcheckpoint database into a well-documented resource, literature-based DbTF curation guidelines utilizing GO controlled vocabularies are required. Literature-curated knowledge sources are generally perceived to be of high-quality. However, due to lack of formalized knowledge representation it is often difficult for a curator to extract accurate information from literature. Therefore there are instances, e.g. in case of protein-protein interaction information curation, where the literature-curated information not necessarily has been of very high quality [195]. In order to curate DbTFs that are verified experimentally in the scientific publications, we formalized DbTF curation guidelines (Paper III). These guidelines describe the use of Gene Ontology controlled vocabularies that are specific for sequence-specific DNA binding factors (e.g. *sequence-specific DNA binding RNAP II transcription factor activity, GO:0000981*), and GOC experimental evidence codes to unambiguously record DbTFs from scientific literature. Similar initiatives have also been implemented for creating GO-annotations of predictive protein signatures from different databases in InterPro database [196], and for the peroxisome proteome in humans [197], where they provide detail protocol for creating specific GO-annotations from literature. Based on our current estimate ~800 DbTFs are experimentally documented in literature. The GOC presently

contains ~200 DbTFs with experimental evidence and we are aiming to add the remaining ~600 DbTFs by the end of 2013. With the implementation of DbTF curation guidelines, we aim to not only improve the quality of the DbTF information content in the TFcheckpoint database but also to enrich the GOC database, which provides an interface for large-scale data analysis and omics data-based work. Our curation guidelines provide the basis both for the ongoing annotation performed in our research group together with GOC, and also for other community researchers who contribute TF-annotations to the GOC (e.g. Saccharomyces Genome Database, The Arabidopsis Information Resource).

To understand the transcriptional regulatory mechanisms that can explain the temporal dynamics of the ~2000 gastrin responsive genes, we utilized the functionality of Network Component Analysis (NCA) [198] (Paper I). NCA calculates the temporal activity of a TF in response to a signal. For this, NCA requires genome-wide time series data, and TF-TG relation information as input. Thus, to obtain the NCA derived transcription factor activity (TFA) profile in response to gastrin, we exploited temporal mRNA profiles of ~2000 gastrin-responsive genes obtained from our 14 hour gastrin treated microarray time-series experiment in AR42J cells, and TF-TG relation matrix collected from the TFactS database [199]. The NCA-derived transcription factor activity profiles of the transcription factors that are part of the CCKR map indicate that the activities of transcription factors EGR1, ELK1, SRF, AP1, ATF2, FOXO1, FOXO3, NFkB are upregulated by gastrin already after 30 minutes, while the activity of transcription factor CREB1 peaks at 2-4 hours, and those of TCF7L2 and NFkB at 10-12 hours. We hypothesized that the signaling components that transmit the gastrin-mediated regulation of the transcription factors, each with their characteristic protein activity profile, will be found both among the

upstream components already present in the CCKR map as well as among other components that have not yet been reported to participate in gastrin responses. For example, our CCKR map suggests that CREB1, which shows delayed activation profile in response to gastrin, has RSK1/2 as upstream regulator. However, using the PPI-based extension of CCKR map, we identified also RPS6KA4 and RPS6KA5 as CREB1- interactors and thus potential CREB1 regulators. Both RPS6KA4/5 have been documented to be associated with delayed activation of CREB1 in other systems [200, 201]. Based on these observations, it is likely that RPS6KA4 and RPS6KA5 also contribute to the gastrin-mediated delayed activation profile of CREB1. Further experimental validation is required to support that these interactions are indeed involved in the gastrin response. Seok et al. [202] utilized a similar strategy to predict transcription factor activity time profiles and they found a correlation in the TG expression and activity of the transcription factors. TF-TG information is one of the inputs on which the NCA approach crucially depends. Sparsity of TF-TG information in the existing resources, e.g. TFactS [199], therefore is a confounding factor leading to uncertainty in NCA-generated estimations of TFA profiles. However, through our text mining efforts extracting TF-TG information from literature [128], we aim to improve the comprehensiveness of these resources.

It is not always straightforward to procure knowledge concerning TF regulators from the knowledge bases. Thus, a platform which provides easy access to the knowledge in such resources would be invaluable for the research community. Keeping this in mind, we created a Gene eXpression Knowledge Base (GeXKB) (Paper IV), which retrieves information on transcription regulators (DbTFs, co-TFs) and their interactors/regulators (e.g. signal transduction components, interacting proteins) from existing resources e.g. GOC [114], IntAct [61], KEGG

[121], and PAZAR [203]. With the help of GeXKB we identified many novel potential regulators of CREB1, NFκB, and TCF7L2 transcription factors that are documented in other systems but not found to be reported in response to gastrin. Our model system to investigate gastrin mediated responses is the AR42J cell. Many of the new regulators proposed by GeXKB are also identified as genes that are expressed in AR42J cells and can therefore participate in gastrin responses. For example, RPS6KA4 and RPS6KA5 (already discussed above) are also returned from GeXKB as potential CREB1 activators [201]. Similarly, GeXKB identifies CYLD and SIRT1 as NFκB repressors [204], and RUNX3 as a repressor of TCF7L2 [205] and an activator of NFκB [206]. All of these regulators appear in the AR42J expressed genes but are not part of the literature curated map of gastrin signaling pathways (CCKR map). This illustrates how GeXKB can suggest potential novel regulators of transcription factors in a given biological system. Investigation of these regulators in adequate experimental systems for gastrin responses can enhance our understanding of gastrin mediated transcription regulation and subsequent cellular outcomes. Similar to GeXKB, today there are several tools that incorporate background knowledge to generate testable hypotheses e.g. PILGRM [207], which combines user's knowledge and literature analysis on microarray genomic data to generate data-driven hypotheses. Hanalyzer [208] and HyQue [209] allow evaluation of hypotheses by integrating background scientific knowledge. Furthermore, Functional Knowledge Transfer (FKT), which is a machine learning algorithm, leverages integration of prior knowledge to generate novel hypotheses for experimental validation [210]. The hypotheses generated from our efforts on integrating the literature-based gastrin mediated signaling network, with experimentally documented large scale protein-protein interaction knowledge [211], are likely to be of high quality and relevance.

Implementation of knowledge obtained from querying GeXKB on our model-driven hypotheses further enhance the relevance of hypotheses generated on gastrin mediated signal transduction. Thus, such hypotheses should be well suited for goal oriented and efficient experimental validation, aimed to deepen our discovery and insights on novel regulators of gastrin mediated cellular responses.

CONCLUSIONS AND FUTURE PERSPECTIVES

We present a model-driven systems biology approach that can contribute to and hinges on knowledge integration and hypothesis generation, furthering a comprehensive understanding of the gastrin-mediated intracellular signaling and cellular decision making events. Reconstruction of gastrin signaling network provides a scaffold for understanding the dynamic behavior of gastrin mediated responses, by integrating genome-wide temporal gene expression data and gene regulatory network. This indicates that in order to collect more accurate knowledge regarding molecular mechanisms underlying cellular responses, it is meaningful to consider a global view in terms of network components' associations and interactions with other components, including upstream regulators. This requires sound knowledge sources (such as GO database, IntAct and http://www.tfcheckpoint.org/) as well as robust strategies to integrate knowledge.

Comprehensive molecular maps such as the CCKR network can serve as valuable starting points for further modeling - both a) quantitative modeling of focused parts of the pathway and b) qualitative approaches with large scale integration of biological background knowledge and new genome-wide experimental data on top of the comprehensive signaling cascades. Signaling networks such as those encoded in our CCKR map can make use of the tool generated by Tiger et al. [212] that uses signal transduction reactions as an input for generating mathematical models. Thus, using such tools together with our comprehensive map of the CCKR-mediated signaling it is now feasible to envisage a roadmap toward dynamical modeling that can enable numerical simulations and generation of functional predictions, which can provide new impetus for research in the field of gastrin and cholecystokinin systems biology.

REFERENCES

1. Snoep, J. and H. Westerhoff, *From isolation to integration, a systems biology approach for building the Silicon Cell*, in *Systems Biology*, L. Alberghina and H.V. Westerhoff, Editors. 2005, Springer Berlin Heidelberg. p. 13-30.
2. Hood, L., et al., *Systems biology at the Institute for Systems Biology.* Briefings in Functional Genomics & Proteomics, 2008. **7**(4): p. 239-248.
3. Kohl, P., et al., *Systems Biology: An Approach.* Clin Pharmacol Ther, 2010. **88**(1): p. 25-33.
4. Kitano, H., *Systems biology: a brief overview.* Science, 2002. **295**(5560): p. 1662-4.
5. Bertalanffy, L.v., *General System Theory*. 1968, New York: George Braziller, Inc.
6. Kholodenko, B., F. Bruggeman, and H. Sauro, *Mechanistic and modular approaches to modeling and inference of cellular regulatory networks Systems Biology*, L. Alberghina and H.V. Westerhoff, Editors. 2005, Springer Berlin / Heidelberg. p. 357-451.
7. Mendoza, E.R., *Systems Biology: Its Past, Present and Potential.* Philippine Science Letters, 2009. **2**(1).
8. Stephanopoulos, I.R.a.G., *Systems Biology:Networks, Models, and Applications*. Vol. II 2006: 74 halftones & line illus. 366.
9. *Centre for Integrative Systems Biology at Imperial College (CISBIC)*. Available from: http://www.doc.ic.ac.uk/bioinformatics/CISB/.
10. Aldridge, B.B., et al., *Physicochemical modelling of cell signalling pathways.* Nature Cell Biology, 2006. **8**(11): p. 1195-1203.
11. Alves, R., F. Antunes, and A. Salvador, *Tools for kinetic modeling of biochemical networks.* Nature Biotechnology, 2006. **24**(6): p. 667-672.
12. Gehlenborg, N., et al., *Visualization of omics data for systems biology.* Nat Methods, 2010. **7**(3 Suppl): p. S56-68.
13. Funahashi, A., et al., *CellDesigner 3.5: A versatile modeling tool for biochemical networks.* Proceedings of the Ieee, 2008. **96**(8): p. 1254-1265.
14. Oda, K., et al., *A comprehensive pathway map of epidermal growth factor receptor signaling.* Mol Syst Biol, 2005. **1**: p. 2005 0010.
15. Caron, E., et al., *A comprehensive map of the mTOR signaling network.* Mol Syst Biol, 2010. **6**: p. 453.
16. Mizuno, S., et al., *AlzPathway: a comprehensive map of signaling pathways of Alzheimer's disease.* BMC Syst Biol, 2012. **6**: p. 52.
17. Albert, R. and R.S. Wang, *Discrete dynamic modeling of cellular signaling networks.* Methods Enzymol, 2009. **467**: p. 281-306.
18. Sun, Z. and R. Albert, *Chapter 10 - Boolean Models of Cellular Signaling Networks*, in *Handbook of Systems Biology*. 2013, Academic Press: San Diego. p. 197-210.
19. Chaouiya, C., *Petri net modelling of biological networks.* Brief Bioinform, 2007. **8**(4): p. 210-9.
20. Kestler, H.A., et al., *Network modeling of signal transduction: establishing the global view.* Bioessays, 2008. **30**(11-12): p. 1110-1125.
21. Kholodenko, B.N., et al., *Quantification of Short Term Signaling by the Epidermal Growth Factor Receptor.* Journal of Biological Chemistry, 1999. **274**(42): p. 30169-30181.
22. Kholodenko, B.N., et al., *Untangling the wires: A strategy to trace functional interactions in signaling and gene networks.* Proceedings of the National Academy of Sciences, 2002. **99**(20): p. 12841-12846.

23. Bruggeman, F.J., et al., *Modular Response Analysis of Cellular Regulatory Networks.* Journal of Theoretical Biology, 2002. **218**(4): p. 507-520.
24. Schilling, M., et al., *Theoretical and experimental analysis links isoform-specific ERK signalling to cell fate decisions.* Mol Syst Biol, 2009. **5**.
25. Orton, R., et al., *Computational modelling of cancerous mutations in the EGFR/ERK signalling pathway.* BMC Systems Biology, 2009. **3**(1): p. 100.
26. Nakakuki, T., et al., *Ligand-specific c-Fos expression emerges from the spatiotemporal control of ErbB network dynamics.* Cell, 2010. **141**(5): p. 884-96.
27. Zhang, R., et al., *Network model of survival signaling in large granular lymphocyte leukemia.* Proc Natl Acad Sci U S A, 2008. **105**(42): p. 16308-13.
28. Schilling, M., et al., *Theoretical and experimental analysis links isoform-specific ERK signalling to cell fate decisions.* Mol Syst Biol, 2009. **5**: p. 334.
29. Bianconi, F., et al., *Computational model of EGFR and IGF1R pathways in lung cancer: a Systems Biology approach for Translational Oncology.* Biotechnol Adv, 2012. **30**(1): p. 142-53.
30. Barabasi, A.L. and Z.N. Oltvai, *Network biology: understanding the cell's functional organization.* Nat Rev Genet, 2004. **5**(2): p. 101-13.
31. Brown, K.S., et al., *The statistical mechanics of complex signaling networks: nerve growth factor signaling.* Phys Biol, 2004. **1**(3-4): p. 184-95.
32. Chuang, H.Y., et al., *Network-based classification of breast cancer metastasis.* Mol Syst Biol, 2007. **3**: p. 140.
33. Schadt, E.E., et al., *An integrative genomics approach to infer causal associations between gene expression and disease.* Nat Genet, 2005. **37**(7): p. 710-7.
34. Barzel, B., A. Sharma, and A.-L. Barabási, *Chapter 9 - Graph Theory Properties of Cellular Networks*, in *Handbook of Systems Biology*. 2013, Academic Press: San Diego. p. 177-193.
35. Schneider, H.C. and T. Klabunde, *Understanding drugs and diseases by systems biology?* Bioorg Med Chem Lett, 2013. **23**(5): p. 1168-76.
36. Bruggeman, F.J. and H.V. Westerhoff, *The nature of systems biology.* Trends in Microbiology, 2007. **15**(1): p. 45-50.
37. van Eunen, K., et al., *Measuring enzyme activities under standardized in vivo-like conditions for systems biology.* FEBS J, 2010. **277**(3): p. 749-60.
38. Rogers, S., *Statistical methods and models for bridging Omics data levels.* Methods Mol Biol, 2011. **719**: p. 133-51.
39. Rosa, B.A., et al., *Computing gene expression data with a knowledge-based gene clustering approach.* Int J Biochem Mol Biol, 2010. **1**(1): p. 51-68.
40. O'Malley, M.A. and J. Dupré, *Fundamental issues in systems biology.* BioEssays, 2005. **27**(12): p. 1270-1276.
41. *Systems Biology: Toward System-level Understanding of Biological Systems -- Kitano 2001.*
42. Luo, C. and Y. Rudy, *A dynamic model of the cardiac ventricular action potential. I. Simulations of ionic currents and concentration changes.* Circulation Research, 1994. **74**(6): p. 1071-1096.
43. ten Tusscher, K.H.W.J., et al., *A model for human ventricular tissue.* American Journal of Physiology - Heart and Circulatory Physiology, 2004. **286**(4): p. H1573-H1589.
44. Plank, G., et al., *Generation of histo-anatomically representative models of the individual heart: tools and application.* Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 2009. **367**(1896): p. 2257-2292.
45. Noble, D., et al., *Resistance of Cardiac Cells to NCX Knockout.* Annals of the New York Academy of Sciences, 2007. **1099**(1): p. 306-309.
46. Noble, D., *The rise of computational biology.* Nat Rev Mol Cell Biol, 2002. **3**(6): p. 459-63.

47. Eisenberg, D., et al., *Protein function in the post-genomic era.* Nature, 2000. **405**(6788): p. 823-6.
48. Copland, J.A., et al., *Sex steroid receptors in skeletal differentiation and epithelial neoplasia: is tissue-specific intervention possible?* Bioessays, 2009. **31**(6): p. 629-41.
49. Rout, M.P., et al., *The yeast nuclear pore complex: composition, architecture, and transport mechanism.* J Cell Biol, 2000. **148**(4): p. 635-51.
50. Uetz, P., et al., *A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae.* Nature, 2000. **403**(6770): p. 623-7.
51. Ito, T., et al., *A comprehensive two-hybrid analysis to explore the yeast protein interactome.* Proc Natl Acad Sci U S A, 2001. **98**(8): p. 4569-74.
52. Gavin, A.C., et al., *Proteome survey reveals modularity of the yeast cell machinery.* Nature, 2006. **440**(7084): p. 631-6.
53. Krogan, N.J., et al., *Global landscape of protein complexes in the yeast Saccharomyces cerevisiae.* Nature, 2006. **440**(7084): p. 637-43.
54. von Mering, C., et al., *Comparative assessment of large-scale data sets of protein-protein interactions.* Nature, 2002. **417**(6887): p. 399-403.
55. Szklarczyk, D., et al., *The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored.* Nucleic Acids Res, 2011. **39**(Database issue): p. D561-8.
56. Roberts, P.M., *Mining literature for systems biology.* Briefings in Bioinformatics, 2006. **7**(4): p. 399-406.
57. Marcotte, E.M. and S.V. Date, *Exploiting big biology: Integrating large-scale biological data for function inference.* Briefings in Bioinformatics, 2001. **2**(4): p. 363-374.
58. Walhout, A.J. and M. Vidal, *Protein interaction maps for model organisms.* Nat Rev Mol Cell Biol, 2001. **2**(1): p. 55-62.
59. De Las Rivas, J. and C. Fontanillo, *Protein-protein interactions essentials: key concepts to building and analyzing interactome networks.* PLoS Comput Biol, 2010. **6**(6): p. e1000807.
60. Keshava Prasad, T.S., et al., *Human Protein Reference Database—2009 update.* Nucleic Acids Research, 2009. **37**(suppl 1): p. D767-D772.
61. Kerrien, S., et al., *The IntAct molecular interaction database in 2012.* Nucleic Acids Research, 2012. **40**(D1): p. D841-D846.
62. Turner, B., et al., *iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence.* Database (Oxford), 2010. **2010**: p. baq023.
63. Sorokin, A.V., E.R. Kim, and L.P. Ovchinnikov, *Proteasome system of protein degradation and processing.* Biochemistry (Mosc), 2009. **74**(13): p. 1411-42.
64. Sweeney, S.E. and G.S. Firestein, *Primer: signal transduction in rheumatic disease--a clinician's guide.* Nat Clin Pract Rheumatol, 2007. **3**(11): p. 651-60.
65. Jordan, J.D., E.M. Landau, and R. Iyengar, *Signaling networks: The origins of cellular multitasking.* Cell, 2000. **103**(2): p. 193-200.
66. Bulyk, M.L. and A.J.M. Walhout, *Chapter 4 - Gene Regulatory Networks*, in *Handbook of Systems Biology*. 2013, Academic Press: San Diego. p. 65-88.
67. Yang, V.W., *Eukaryotic transcription factors: identification, characterization and functions.* J Nutr, 1998. **128**(11): p. 2045-51.
68. Klepper, K. and F. Drablos, *MotifLab: a tools and data integration workbench for motif discovery and regulatory sequence analysis.* Bmc Bioinformatics, 2013. **14**.
69. Fu, Y. and W. Xiao, *Study of Transcriptional Regulation Using a Reporter Gene Assay*, in *Yeast Protocol*, W. Xiao, Editor. 2006, Humana Press. p. 257-264.
70. Hannon, G.J., *RNA interference.* Nature, 2002. **418**(6894): p. 244-251.

71. De-Leon, S.B.T. and E.H. Davidson, *Gene regulation: Gene control network in development.* Annual Review of Biophysics and Biomolecular Structure, 2007. **36**: p. 191-212.
72. Davidson, E.H., et al., *A genomic regulatory network for development.* Science, 2002. **295**(5560): p. 1669-1678.
73. Gerstein, M.B., et al., *Architecture of the human regulatory network derived from ENCODE data.* Nature, 2012. **489**(7414): p. 91-100.
74. Rosenbloom, K.R., et al., *ENCODE whole-genome data in the UCSC Genome Browser: update 2012.* Nucleic Acids Research, 2012. **40**(D1): p. D912-D917.
75. Wang, J., et al., *Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors.* Genome Research, 2012. **22**(9): p. 1798-1812.
76. Gerstein, M.B., et al., *Architecture of the human regulatory network derived from ENCODE data.* Nature, 2012. **489**(7414): p. 91-100.
77. Fleischmann, R.D., et al., *Whole-Genome Random Sequencing and Assembly of Haemophilus-Influenzae Rd.* Science, 1995. **269**(5223): p. 496-512.
78. Collins, F.S., et al., *Finishing the euchromatic sequence of the human genome.* Nature, 2004. **431**(7011): p. 931-945.
79. Harel, A., et al., *Omics Data Management and Annotation*, in *Bioinformatics for Omics Data*, B. Mayer, Editor. 2011, Humana Press. p. 71-96.
80. Joyce, A.R. and B.O. Palsson, *The model organism as a system: integrating 'omics' data sets.* Nat Rev Mol Cell Biol, 2006. **7**(3): p. 198-210.
81. Taylor, C.F., et al., *Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project.* Nature Biotechnology, 2008. **26**(8): p. 889-896.
82. Smith, B., et al., *The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration.* Nature Biotechnology, 2007. **25**(11): p. 1251-1255.
83. Jones, A.R., et al., *The Functional Genomics Experiment model (FuGE): an extensible framework for standards in functional genomics.* Nature Biotechnology, 2007. **25**(10): p. 1127-1133.
84. Field, D., et al., *The minimum information about a genome sequence (MIGS) specification.* Nature Biotechnology, 2008. **26**(5): p. 541-547.
85. Brazma, A., et al., *Minimum information about a microarray experiment (MIAME) - toward standards for microarray data.* Nature Genetics, 2001. **29**(4): p. 365-371.
86. Orchard, S., et al., *Further steps towards data standardisation: the Proteomic Standards Initiative HUPO 3(rd) annual congress, Beijing 25-27(th) October, 2004.* Proteomics, 2005. **5**(2): p. 337-9.
87. Taylor, C.F., et al., *The minimum information about a proteomics experiment (MIAPE).* Nature Biotechnology, 2007. **25**(8): p. 887-893.
88. Lindon, J.C., et al., *Summary recommendations for standardization and reporting of metabolic analyses.* Nat Biotechnol, 2005. **23**(7): p. 833-8.
89. Rebhan, M., et al., *GeneCards: integrating information about genes, proteins and diseases.* Trends Genet, 1997. **13**(4): p. 163.
90. Rebhan, M., et al., *GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support.* Bioinformatics, 1998. **14**(8): p. 656-64.
91. Safran, M., et al., *GeneCards 2002: towards a complete, object-oriented, human gene compendium.* Bioinformatics, 2002. **18**(11): p. 1542-3.
92. Chalifa-Caspi, V., et al., *GeneAnnot: interfacing GeneCards with high-throughput gene expression compendia.* Brief Bioinform, 2003. **4**(4): p. 349-60.
93. Safran, M., et al., *GeneCards Version 3: the human gene integrator.* Database (Oxford), 2010. **2010**: p. baq020.

94. Haider, S., et al., *BioMart Central Portal--unified access to biological data.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W23-7.
95. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nat Protoc, 2009. **4**(1): p. 44-57.
96. Baitaluk, M., et al., *IntegromeDB: an integrated system and biological search engine.* BMC Genomics, 2012. **13**: p. 35.
97. Janes, K.A. and M.B. Yaffe, *Data-driven modelling of signal-transduction networks.* Nature Reviews Molecular Cell Biology, 2006. **7**(11): p. 820-828.
98. Bansal, M., et al., *How to infer gene networks from expression profiles.* Mol Syst Biol, 2007. **3**.
99. Lee, W.-P. and W.-S. Tzou, *Computational methods for discovering gene networks from expression data.* Briefings in Bioinformatics, 2009. **10**(4): p. 408-423.
100. Bader, G.D., M.P. Cary, and C. Sander, *Pathguide: a Pathway Resource List.* Nucleic Acids Research. **34**(suppl 1): p. D504-D506.
101. Boutet, E., et al., *UniProtKB/Swiss-Prot.* Methods Mol Biol, 2007. **406**: p. 89-112.
102. Rose, P.W., et al., *The RCSB Protein Data Bank: new resources for research and education.* Nucleic Acids Res, 2013. **41**(Database issue): p. D475-82.
103. Li, C., et al., *BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models.* BMC Syst Biol, 2010. **4**: p. 92.
104. Croft, D., et al., *Reactome: a database of reactions, pathways and biological processes.* Nucleic Acids Res, 2011. **39**(Database issue): p. D691-7.
105. Mi, H., A. Muruganujan, and P.D. Thomas, *PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees.* Nucleic Acids Res, 2013. **41**(Database issue): p. D377-86.
106. Pagani, I., et al., *The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata.* Nucleic Acids Res, 2012. **40**(Database issue): p. D571-9.
107. Benson, D.A., et al., *GenBank.* Nucleic Acids Research, 2011. **39**(suppl 1): p. D32-D37.
108. Barrett, T. and R. Edgar, *Gene expression omnibus: microarray data storage, submission, retrieval, and analysis.* Methods Enzymol, 2006. **411**: p. 352-69.
109. Parkinson, H., et al., *ArrayExpress update—an archive of microarray and high-throughput sequencing-based functional genomics experiments.* Nucleic Acids Research, 2011. **39**(suppl 1): p. D1002-D1004.
110. de Souza, N., *The ENCODE project.* Nat Methods, 2012. **9**(11): p. 1046.
111. Bolser, D.M., et al., *MetaBase--the wiki-database of biological databases.* Nucleic Acids Res, 2012. **40**(Database issue): p. D1250-4.
112. Bader, G.D., M.P. Cary, and C. Sander, *Pathguide: a pathway resource list.* Nucleic Acids Res, 2006. **34**(Database issue): p. D504-6.
113. Gruber, T.R., *Toward principles for the design of ontologies used for knowledge sharing?* International Journal of Human-Computer Studies, 1995. **43**(5–6): p. 907-928.
114. Gene Ontology, C., *Gene Ontology annotations and resources.* Nucleic Acids Res, 2013. **41**(Database issue): p. D530-5.
115. Antezana, E., et al., *The Cell Cycle Ontology: an application ontology for the representation and integrated analysis of the cell cycle process.* Genome Biology, 2009. **10**(5).
116. Smith, B., *Basic concepts of formal ontology.* Formal Ontology in Information Systems, 1998. **46**: p. 19-28.
117. Whetzel, P.L., et al., *BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies*

in software applications. Nucleic Acids Res, 2011. **39**(Web Server issue): p. W541-5.
118. *OWL Web Ontology Language*. Available from: http://www.w3.org/TR/owl-features/.
119. *Resource Description Framework*. Available from: http://www.w3.org/RDF/.
120. Rocca-Serra, P., et al., *ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level.* Bioinformatics, 2010. **26**(18): p. 2354-6.
121. Kanehisa, M., et al., *KEGG for integration and interpretation of large-scale molecular data sets.* Nucleic Acids Res, 2012. **40**(Database issue): p. D109-14.
122. Maglott, D., et al., *Entrez Gene: gene-centered information at NCBI.* Nucleic Acids Res, 2011. **39**(Database issue): p. D52-7.
123. Flicek, P., et al., *Ensembl 2013*. Nucleic Acids Res, 2013. **41**(Database issue): p. D48-55.
124. Hill, D.P., et al., *Gene Ontology annotations: what they mean and where they come from.* BMC Bioinformatics, 2008. **9 Suppl 5**: p. S2.
125. Cusick, M.E., et al., *Literature-curated protein interaction datasets.* Nat Meth, 2009. **6**(1): p. 39-46.
126. Kemper, B., et al., *PathText: a text mining integrator for biological pathway visualizations*. Bioinformatics, 2010. **26**(12): p. i374-81.
127. Ananiadou, S., et al., *Event extraction for systems biology by text mining the literature.* Trends in Biotechnology, 2010. **28**(7): p. 381-390.
128. Florian Leitner, M.K., Sushil Tripathi, Martin Kuiper, Astrid Lægreid, and Alfonso Valencia, *Mining cis-Regulatory Transcription Networks from Literature*, in *ISMB/ECCB SIG*. 2013.
129. Roldan-Garcia Mdel, M., et al., *KA-SB: from data integration to large scale reasoning.* BMC Bioinformatics, 2009. **10 Suppl 10**: p. S5.
130. Goble, C. and R. Stevens, *State of the nation in data integration for bioinformatics.* Journal of Biomedical Informatics, 2008. **41**(5): p. 687-693.
131. Antezana, E., et al., *BioGateway: a semantic systems biology tool for the life sciences.* BMC Bioinformatics, 2009. **10 Suppl 10**: p. S11.
132. Federhen, S., *The NCBI Taxonomy database.* Nucleic Acids Res, 2012. **40**(Database issue): p. D136-43.
133. *Gene Ontology Annotation Files*. Available from: http://www.geneontology.org/ontology/.
134. *SPARQL Query Language for RDF*. Available from: http://www.w3.org/TR/rdf-sparql-query/.
135. Shannon, P.T., et al., *The Gaggle: an open-source software system for integrating bioinformatics software and data sources.* BMC Bioinformatics, 2006. **7**: p. 176.
136. Smith, R.N., et al., *InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data.* Bioinformatics, 2012. **28**(23): p. 3163-5.
137. Watson, S.A., et al., *Gastrin - active participant or bystander in gastric carcinogenesis?* Nature Reviews Cancer, 2006. **6**(12): p. 936-946.
138. Dufresne, M., C. Seva, and D. Fourmy, *Cholecystokinin and gastrin receptors.* Physiological Reviews, 2006. **86**(3): p. 805-847.
139. Noble, F., et al., *International Union of Pharmacology. XXI. Structure, distribution, and functions of cholecystokinin receptors.* Pharmacol Rev, 1999. **51**(4): p. 745-81.
140. Watson, S.A., et al., *Gastrin - active participant or bystander in gastric carcinogenesis?* Nat Rev Cancer, 2006. **6**(12): p. 936-46.
141. Wang, T.C., et al., *Synergistic interaction between hypergastrinemia and Helicobacter infection in a mouse model of gastric cancer.* Gastroenterology, 2000. **118**(1): p. 36-47.

142. Zavros, Y., et al., *Chronic gastritis in the hypochlorhydric gastrin-deficient mouse progresses to adenocarcinoma.* Oncogene, 2005. **24**(14): p. 2354-2366.
143. Beales, I., et al., *Effect of Helicobacter pylori products and recombinant cytokines on gastrin release from cultured canine G cells.* Gastroenterology, 1997. **113**(2): p. 465-471.
144. Webb, P.M., et al., *Gastric cancer and Helicobacter pylori: a combined analysis of 12 case control studies nested within prospective cohorts.* Gut, 2001. **49**(3): p. 347-353.
145. Grabowska, A.M. and S.A. Watson, *Role of gastrin peptides in carcinogenesis.* Cancer Lett, 2007. **257**(1): p. 1-15.
146. Cui, G.L., et al., *Gastrin-induced apoptosis contributes to carcinogenesis in the stomach.* Laboratory Investigation, 2006. **86**(10): p. 1037-1051.
147. Cayrol, C., et al., *Cholecystokinin-2 receptor modulates cell adhesion through beta 1-integrin in human pancreatic cancer cells.* Oncogene, 2006. **25**(32): p. 4421-4428.
148. Rao, J.N. and J.Y. Wang, in *Regulation of Gastrointestinal Mucosal Growth*. 2010: San Rafael (CA).
149. Prinz, C., et al., *Gastrin effects on isolated rat enterochromaffin-like cells in primary culture.* Am J Physiol, 1994. **267**(4 Pt 1): p. G663-75.
150. Tielemans, Y., et al., *Self-replication of enterochromaffin-like cells in the mouse stomach.* Digestion, 1990. **45**(3): p. 138-46.
151. Koh, T.J. and D. Chen, *Gastrin as a growth factor in the gastrointestinal tract.* Regulatory Peptides, 2000. **93**(1–3): p. 37-44.
152. Blackmore, M. and B.H. Hirst, *Autocrine stimulation of growth of AR4-2J rat pancreatic tumour cells by gastrin.* Br J Cancer, 1992. **66**(1): p. 32-8.
153. Ishizuka, J., et al., *The effect of gastrin on growth of human stomach cancer cells.* Ann Surg, 1992. **215**(5): p. 528-34.
154. Dehez, S., et al., *Gastrin-induced DNA synthesis requires p38-MAPK activation via PKC/Ca(2+) and Src-dependent mechanisms.* FEBS Lett, 2001. **496**(1): p. 25-30.
155. Dehez, S., et al., *c-Jun NH(2)-terminal kinase pathway in growth-promoting effect of the G protein-coupled receptor cholecystokinin B receptor: a protein kinase C/Src-dependent-mechanism.* Cell Growth Differ, 2002. **13**(8): p. 375-85.
156. Stepan, V.M., et al., *Cell type-specific requirement of the MAPK pathway for the growth factor action of gastrin.* Am J Physiol, 1999. **276**(6 Pt 1): p. G1363-72.
157. Dabrowski, A., et al., *Stimulation of both CCK-A and CCK-B receptors activates MAP kinases in AR42J and receptor-transfected CHO cells.* Digestion, 1997. **58**(4): p. 361-7.
158. He, H. and G.S. Baldwin, *Rho GTPases and p21-activated kinase in the regulation of proliferation and apoptosis by gastrins.* Int J Biochem Cell Biol, 2008. **40**(10): p. 2018-22.
159. Stepan, V., et al., *Role of small GTP binding proteins in the growth-promoting and antiapoptotic actions of gastrin.* Am J Physiol Gastrointest Liver Physiol, 2004. **287**(3): p. G715-25.
160. Pradeep, A., et al., *Gastrin-mediated activation of cyclin D1 transcription involves beta-catenin and CREB pathways in gastric cancer cells.* Oncogene, 2004. **23**(20): p. 3689-99.
161. Steigedal, T.S., et al., *Inducible cAMP early repressor suppresses gastrin-mediated activation of cyclin D1 and c-fos gene expression.* Am J Physiol Gastrointest Liver Physiol, 2007. **292**(4): p. G1062-9.
162. Muerkoster, S., et al., *Gastrin suppresses growth of CCK2 receptor expressing colon cancer cells by inducing apoptosis in vitro and in vivo.* Gastroenterology, 2005. **129**(3): p. 952-968.

163. Hall, P.A., et al., *Regulation of cell number in the mammalian gastrointestinal tract: the importance of apoptosis.* J Cell Sci, 1994. **107 (Pt 12)**: p. 3569-77.
164. Przemeck, S.M., et al., *Hypergastrinemia increases gastric epithelial susceptibility to apoptosis.* Regul Pept, 2008. **146**(1-3): p. 147-56.
165. Zhou, Q., et al., *Role of REG Ialpha in gastric carcinogenesis: Gastrin-associated proliferative and anti-apoptotic activities.* Mol Med Rep, 2010. **3**(6): p. 999-1005.
166. Todisco, A., et al., *Molecular mechanisms for the antiapoptotic action of gastrin.* Am J Physiol Gastrointest Liver Physiol, 2001. **280**(2): p. G298-307.
167. Cui, G., et al., *Gastrin-induced apoptosis contributes to carcinogenesis in the stomach.* Lab Invest, 2006. **86**(10): p. 1037-51.
168. Ramamoorthy, S., V. Stepan, and A. Todisco, *Intracellular mechanisms mediating the anti-apoptotic action of gastrin.* Biochem Biophys Res Commun, 2004. **323**(1): p. 44-8.
169. Fjeldbo, C.S., et al., *Gastrin upregulates the prosurvival factor secretory clusterin in adenocarcinoma cells and in oxyntic mucosa of hypergastrinemic rats.* Am J Physiol Gastrointest Liver Physiol, 2012. **302**(1): p. G21-33.
170. Pritchard, D.M., et al., *Gastrin increases mcl-1 expression in type I gastric carcinoid tumors and a gastric epithelial cell line that expresses the CCK-2 receptor.* Am J Physiol Gastrointest Liver Physiol, 2008. **295**(4): p. G798-805.
171. Konturek, P.C., et al., *Influence of gastrin on the expression of cyclooxygenase-2, hepatocyte growth factor and apoptosis-related proteins in gastric epithelial cells.* J Physiol Pharmacol, 2003. **54**(1): p. 17-32.
172. Ferrand, A., et al., *Involvement of JAK2 upstream of the PI 3-kinase in cell-cell adhesion regulation by gastrin.* Exp Cell Res, 2004. **301**(2): p. 128-38.
173. Noble, P.J., et al., *Stimulation of gastrin-CCKB receptor promotes migration of gastric AGS cells via multiple paracrine pathways.* Am J Physiol Gastrointest Liver Physiol, 2003. **284**(1): p. G75-84.
174. Mishra, P., et al., *Mixed lineage kinase-3/JNK1 axis promotes migration of human gastric cancer cells following gastrin stimulation.* Mol Endocrinol, 2010. **24**(3): p. 598-607.
175. Yu, H.G., et al., *p190RhoGEF (Rgnef) promotes colon carcinoma tumor progression via interaction with focal adhesion kinase.* Cancer Res, 2011. **71**(2): p. 360-70.
176. Ding, J., et al., *[Effect of gastrin on invasiveness of human colon cancer cells].* Zhonghua Zhong Liu Za Zhi, 2005. **27**(4): p. 213-5.
177. Wroblewski, L.E., et al., *Gastrin-stimulated gastric epithelial cell invasion: the role and mechanism of increased matrix metalloproteinase 9 expression.* Biochem J, 2002. **365**(Pt 3): p. 873-9.
178. Mishra, P., et al., *Glycogen Synthase Kinase-3beta regulates Snail and beta-catenin during gastrin-induced migration of gastric cancer cells.* J Mol Signal, 2010. **5**: p. 9.
179. Norsett, K.G., et al., *Gastrin stimulates expression of plasminogen activator inhibitor-1 in gastric epithelial cells.* Am J Physiol Gastrointest Liver Physiol, 2011. **301**(3): p. G446-53.
180. Almeida-Vega, S., et al., *Gastrin activates paracrine networks leading to induction of PAI-2 via MAZ and ASC-1.* Am J Physiol Gastrointest Liver Physiol, 2009. **296**(2): p. G414-23.
181. Linn-Karina M Selvik, C.S.F., Arnar Flatberg, Tonje S. Steigedal, Kristine Misund, Endre Anderssen; Berit Doseth, Mette Langaas, Sushil Tripathi, Vidar Beisvag, Astrid Lægreid, Liv Thommesen and Torunn Bruland, *The duration of gastrin treatment affects global gene expression and molecular responses involved in ER stress and anti-apoptosis, revised manuscript awaiting final editorial decision*, in *BMC Genomics*. 2013.

182. Ideker, T., J. Dutkowski, and L. Hood, *Boosting Signal-to-Noise in Complex Biology: Prior Knowledge Is Power.* Cell, 2011. **144**(6): p. 860-863.
183. Arighi, C.N., et al., *An overview of the BioCreative 2012 Workshop Track III: interactive text mining task.* Database (Oxford), 2013. **2013**: p. bas056.
184. Papin, J.A., et al., *Reconstruction of cellular signalling networks and analysis of their properties.* Nat Rev Mol Cell Biol, 2005. **6**(2): p. 99-111.
185. Oda, K. and H. Kitano, *A comprehensive map of the toll-like receptor signaling network.* Mol Syst Biol, 2006. **2**: p. 2006 0015.
186. Fink, M.Y., et al., *Research Resource: Gonadotropin-Releasing Hormone Receptor-Mediated Signaling Network in L beta T2 Cells: A Pathway-Based Web-Accessible Knowledgebase.* Molecular Endocrinology, 2010. **24**(9): p. 1863-1871.
187. Gloaguen, P., et al., *Mapping the follicle-stimulating hormone-induced signalling networks.* Frontiers in Endocrinology, 2011. **2**.
188. Matsuoka, Y., et al., *Payao: a community platform for SBML pathway model curation.* Bioinformatics, 2010. **26**(10): p. 1381-3.
189. Fink, M.Y., et al., *Research resource: Gonadotropin-releasing hormone receptor-mediated signaling network in LbetaT2 cells: a pathway-based web-accessible knowledgebase.* Mol Endocrinol, 2010. **24**(9): p. 1863-71.
190. Perissi, V., et al., *Deconstructing repression: evolving models of co-repressor action.* Nat Rev Genet, 2010. **11**(2): p. 109-23.
191. Weake, V.M. and J.L. Workman, *Inducible gene expression: diverse regulatory mechanisms.* Nat Rev Genet, 2010. **11**(6): p. 426-37.
192. Fulton, D.L., et al., *TFCat: the curated catalog of mouse and human transcription factors.* Genome Biol, 2009. **10**(3): p. R29.
193. Zhang, H.M., et al., *AnimalTFDB: a comprehensive animal transcription factor database.* Nucleic Acids Res, 2012. **40**(Database issue): p. D144-9.
194. Yusuf, D., et al., *The transcription factor encyclopedia.* Genome Biol, 2012. **13**(3): p. R24.
195. Cusick, M.E., et al., *Literature-curated protein interaction datasets.* Nat Methods, 2009. **6**(1): p. 39-46.
196. Burge, S., et al., *Manual GO annotation of predictive protein signatures: the InterPro approach to GO curation.* Database (Oxford), 2012. **2012**: p. bar068.
197. Mutowo-Meullenet, P., et al., *Use of Gene Ontology Annotation to understand the peroxisome proteome in humans.* Database-the Journal of Biological Databases and Curation, 2013.
198. Liao, J.C., et al., *Network component analysis: Reconstruction of regulatory signals in biological systems.* Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(26): p. 15522-15527.
199. Essaghir, A., et al., *Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data.* Nucleic Acids Research, 2010. **38**(11).
200. Wu, G.Y., K. Deisseroth, and R.W. Tsien, *Activity-dependent CREB phosphorylation: convergence of a fast, sensitive calmodulin kinase pathway and a slow, less sensitive mitogen-activated protein kinase pathway.* Proc Natl Acad Sci U S A, 2001. **98**(5): p. 2808-13.
201. Delghandi, M.P., M. Johannessen, and U. Moens, *The cAMP signalling pathway activates CREB through PKA, p38 and MSK1 in NIH 3T3 cells.* Cellular Signalling, 2005. **17**(11): p. 1343-1351.
202. Seok, J., et al., *A dynamic network of transcription in LPS-treated human subjects.* BMC Syst Biol, 2009. **3**: p. 78.
203. Portales-Casamar, E., et al., *The PAZAR database of gene regulatory information coupled to the ORCA toolkit for the study of regulatory sequences.* Nucleic Acids Res, 2009. **37**(Database issue): p. D54-60.

204. Sun, S.C., *CYLD: a tumor suppressor deubiquitinase regulating NF-kappaB activation and diverse biological processes.* Cell Death Differ, 2010. **17**(1): p. 25-34.
205. Ito, K., et al., *RUNX3 attenuates beta-catenin/T cell factors in intestinal tumorigenesis.* Cancer Cell, 2008. **14**(3): p. 226-37.
206. Lim, B., et al., *Increased Genetic Susceptibility to Intestinal-Type Gastric Cancer Is Associated With Increased Activity of the RUNX3 Distal Promoter.* Cancer, 2011. **117**(22): p. 5161-5171.
207. Greene, C.S. and O.G. Troyanskaya, *PILGRM: an interactive data-driven discovery platform for expert biologists.* Nucleic Acids Res, 2011. **39**(Web Server issue): p. W368-74.
208. Leach, S.M., et al., *Biomedical Discovery Acceleration, with Applications to Craniofacial Development.* Plos Computational Biology, 2009. **5**(3).
209. Callahan, A., M. Dumontier, and N.H. Shah, *HyQue: evaluating hypotheses using Semantic Web technologies.* J Biomed Semantics, 2011. **2 Suppl 2**: p. S3.
210. Park, C.Y., et al., *Functional knowledge transfer for high-accuracy prediction of under-studied biological processes.* PLoS Comput Biol, 2013. **9**(3): p. e1002957.
211. Glaab, E., et al., *Extending pathways and processes using molecular interaction networks to analyse cancer genome data.* Bmc Bioinformatics, 2010. **11**.
212. Tiger, C.F., et al., *A framework for mapping, visualisation and automatic model creation of signal-transduction networks.* Mol Syst Biol, 2012. **8**: p. 578.

PAPER I-IV

Paper I

The Gastrin and Cholecystokinin Receptors mediated signaling network: A scaffold for data analysis and new hypotheses on regulatory mechanisms

Sushil Tripathi[1], Åsmund Flobak[1], Konika Chawla[2], Anaïs Baudot[3], Jayavelu Naresh Doni[4], Nadav Skjøndal-Bar[4], Torunn Bruland[1], Liv Thommesen[1, 5], Martin Kuiper[2] and Astrid Lægreid[1*]

[1] Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology (NTNU), N-7489 Trondheim, Norway.

[2] Department of Biology, Norwegian University of Science and Technology (NTNU), N-7491 Trondheim, Norway.

[3] IML, Institut de Mathématiques de Luminy Campus de Luminy, Case 907, 13288 MARSEILLE Cedex 9, France.

[4] Department of Chemical Engineering, Norwegian University of Science and Technology (NTNU), N-7491 Trondheim, Norway.

[5] Department of Technology, Sør-Trøndelag University College, N-7004 Trondheim, Norway.

*Corresponding author. Institute of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology (NTNU), N-7489 Trondheim, Norway. Tel.: +4772825323 Fax : +4772571463; E-mail : astrid.lagreid@ntnu.no

Email addresses:
ST: sushil.tripathi@ntnu.no
ÅF: asmund.flobak@ntnu.no
KC: konika.chawla@ntnu.no
AB: anaisbaudot@gmail.com
JND: jayavelu.n.doni@ntnu.no
NSB: nadi.bar@chemeng.ntnu.no
TB: torunn.bruland@ntnu.no
LT: liv.thommesen@ntnu.no
MK: martin.kuiper@ntnu.no
AL: astrid.lagreid@ntnu.no

Abstract

Background

The gastrointestinal peptide hormones gastrin and cholecystokinin exert their biological functions via their receptors cholecystokinin receptor (CCKR) 1 and 2. Gastrin is a central regulator of gastric acid secretion and growth and differentiation of gastric and colonic mucosa, and is suggested to be pro-carcinogenic. Cholecystokinin is implicated in digestion, appetite control and body weight regulation and may play a role in several digestive disorders. A comprehensive map of gastrin and cholecystokinin receptors mediated signaling cascades that supports systems level studies of these hormones does not exist.

Results

We built a literature-curated map of cholecystokinin receptors 1 and -2 signaling pathways mediating biological responses to gastrin and cholecystokinin. Computational decomposition of the cholecystokinin receptor signaling map into sub-networks revealed 18 modules, representing higher level structures of the signaling network and offering an enhanced understanding of intracellular processes involved in cellular decisions leading towards proliferation, migration and apoptosis. Extension with large scale protein-protein interaction data yielded more than 4000 proteins directly interacting with signaling map components. Topological analyses allowed the prediction of new candidate regulators of gastrin and cholecystokinin signaling based on their ability to increase the compactness of the network. The CCKR model was constructed using the CellDesigner software (http://www.celldesigner.org/) and is freely available together with the module and protein interaction knowledge data.

Conclusion

We here demonstrate how the literature-based CCKR signaling model, its protein interactor extensions and genome-scale time series transcriptome data can be integrated to generate new hypotheses on temporal regulation of molecular mechanisms underlying dynamic cellular processes.

Keywords: cholecystokinin receptor/comprehensive map/modules/transcription factor activities/ analysis/protein-protein interaction

---

Background

Gastrin and cholecystokinin (CCK) are gastrointestinal peptide hormones that share a common C-terminal pentapeptide amide and are produced primarily in G cells of the gastric antrum and I cells of the small intestine, respectively [1]. Gastrin is the central regulator of gastric acid secretion and also regulates growth and differentiation of gastric and colonic mucosa [2]. CCK is involved in physiological processes such as digestion, appetite control and body weight regulation [3]. The scientific interest in these hormones is further strengthened by their roles in several diseases. CCK has been implicated in acute pancreatitis [4-6], obesity [7, 8], irritable bowel syndrome [9] and gallbladder disease [10, 11]. Gastrin is suggested to be pro-carcinogenic, affecting proliferation, angiogenesis and apoptosis [2]; and a co-risk factor for gastric carcinogenesis and atrophy in *Helicobacter pylori* infection [12, 13].

Gastrin and CCK impinge on cellular functions by binding to two different G protein-coupled receptors, CCK1R and CCK2R, located on multiple cell types in peripheral organs, such as the gastrointestine, the pancreas, and the gall bladder [14].

Gastrin has a strong preference for CCK2R, while CCK can activate both receptors with similar affinity [10]. Most cell types responsive to one or both peptide hormones express only one CCK-receptor variant. However, some cells express both CCK1R and CCK2R, including both normal and cancer cells in intact organisms as well as model cell lines such as the rat pancreatic acinar cell derived cell-line AR42J [15].

Today, no comprehensive map of gastrin and CCK signaling exists. A conceptual model that presents known signaling mechanisms for both CCK1R and CCK2R in one framework would be of significant support in systems level studies addressing the differential or combined effects of these hormones. The Nuclear Receptor Signaling Atlas (NURSA) (http://www.nursa.org/) is one of the resources which provides high quality curated knowledge about signaling components and integrate this with genome scale data [16, 17]. Similar to NURSA initiative, the signaling pathway model presented here synthesizes published molecular mechanisms on both specific and shared CCK1R and CCK2R signaling and as such provides a foundation for network-based analyses targeting the identification of signaling hubs, modular structure and regulatory principles. Detailed signaling networks provide a scaffold for understanding cellular aberrations resulting from disease and for identification of central mechanistic disease modules thus enabling identification of therapeutic chemicals that are able to perturb disease module activity. Strategies that build on a systems level understanding have among others allowed drug-induced rewiring of the 'state' of oncogenic signaling networks to maximize the susceptibility to anticancer drugs [18]. Similarly, the resources presented here could be instrumental in the identification of key targets for diseases involving gastrin or CCK.

In the past decade several manually reconstructed comprehensive networks of signaling events have been published [19-27] each of them providing considerable impetus to a systems understanding of signaling mechanisms. The present work extends these efforts to the domain of cholecystokinin receptors (CCKR) signaling by providing a comprehensive literature-based CCKR signaling network model that comprises 519 molecular species and 424 reactions. Segmentation of this vast signaling network into modules using the BiNoM tool [28] resulted in 18 modules, each of which represents a distinct molecular signaling sub-network that interacts with other modules to elicit the differential intracellular signaling responses by gastrin and/or CCK. We then used the CCKR model as a scaffold for further data integration and identified ~4000 proteins directly interacting with at least one of the CCKR model proteins. Further topological analyses based on network connectivity and compactness criteria [29] revealed ~100 tightly connected protein interactors that should be highly ranked as potential regulators of the CCKR mediated signaling network – with central roles in either individual modules, or in the co-ordination of several modules. Finally, we present use cases that demonstrate how the CCKR model and its PPI extensions can provide interesting hypotheses for further refinement of molecular mechanisms governing CCKR intracellular signaling and for improved understanding the dynamics of transcription regulation.

Results

CCKR map building and availability

We present a comprehensive map of the CCKR signaling network (Figure 1). The CCKR model represents a manual assembly of information retrieved from more than 250 scientific publications, and encompasses 214 unique proteins and their relationships to complexes and genes, described by reactions such as state transitions, transport, and heterodimer associations/dissociations (Table 1). The model was written with the use of CellDesigner 4.2 network editor

(http://www.celldesigner.org/) [30] and is available in the data exchange format SBML (Additional file 1). In addition, we have made available a Payao [31] web version, for a full view of all details concerning components, reactions, metadata and references at http://sblab.celldesigner.org:18080/Payao11/bin/ (model name: comprehensive_CCKR_map). Some pathway members, reactions and modules of the CCKR signaling mechanisms can be found in the Reactome [32] and KEGG [33] databases. However, the CCKR map constitutes a much more comprehensive and integrated model providing detailed signaling reactions linking the receptors (CCK1R and CCK2R) all the way down to regulated genes and cellular responses; therefore it represents a significant increase in encoded signaling information as the knowledge currently in Reactome covers less than 5% of the pathway details presented here.

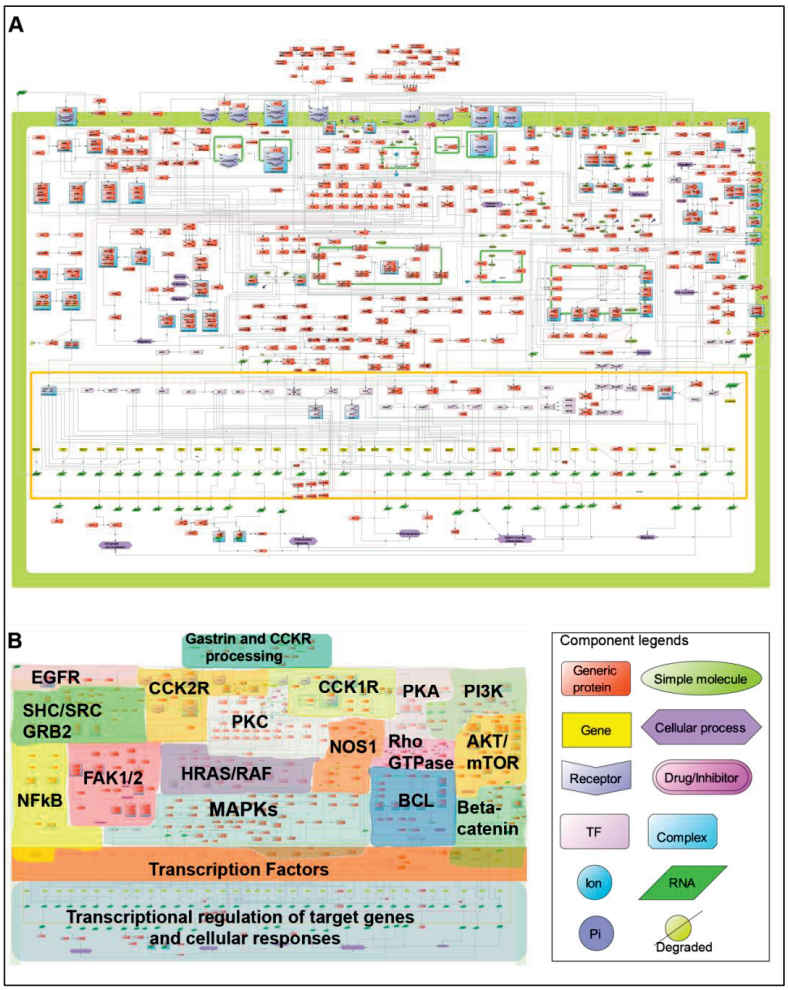


Figure 1. CCKR signaling map (CellDesigner™ 4.2). A. Literature curated comprehensive map of the CCK1R, CCK2R mediated signaling pathways comprising 519 species and 424 reactions (Table 1for further details). The graphical representation is also available as SBML file (Additional file 1) **B.** Navigation map to track components and signaling cascades in the detailed map.

Table 1. Statistical overview of the CCKR map

Species		Reaction	
Category	**Number (519)**	**Category**	**Number (424)**
Proteins	304 (214 unique)	Heterodimer associations and dissociations	43
Complexes	61	State transitions	190
Genes	36	Transports	59
RNAs	63 (35 unique)	Transcriptions and translations	41
Other[a]	55	Other[b]	91

[a]Simple molecules, phenotype, degraded products, ions, drugs, unknown molecules
[b]Known transition omitted, truncation, unknown transitions, unknown negative influence, positive influence.

Signaling pathways shared by CCK1R and CCK2R

Signaling mechanisms downstream of both CCK1R and CCK2R include the trimeric guanine nucleotide binding protein (G protein) alpha q protein ($G\alpha_q$), Protein kinase C (PKC) dependent phosphorylation of adaptor protein Src homology 2 domain-containing-transforming protein C (SHC) and its association with (GRB2)/ (SOS1) leading to activation of the HRAS/RAF/MAPK1/3 cascade. Shared are also pathways involving MAP3K11 mediated MAPK8, -9 and -10 and p38MAPK (MAPK14), PRKD1 (in PKC signaling), PI3K, AKT1, Focal adhesion kinase 1 (FAK1), and Rho GTPase. Both receptors activate PKC isoforms PKCα, -δ, -ε, -θ and -ζ [34-37]. Transcription factors (TFs) reported downstream of both CCKR1 and CCKR2 receptors include NFκB, CREB1, ELK1 and AP1.

Signaling pathways specific for CCK1R

Two trimeric G-proteins appear to be regulated only by CCK1R. One is G alpha S ($G\alpha_s$) [38], which leads to Protein kinase A (PKA) activation via adenylate cyclase catalyzed cAMP production. The other is G alpha 13 ($G\alpha_{13}$) [39], involved in downstream activation of Ras homolog family member A (RHOA) [40]. The nitric oxide synthase (NOS1) signal transduction pathway downstream of CCK1R [41, 42], regulates Ca^{2+} signaling pathways by opening ryanodine receptors and two-pore channels that release calcium from endoplasmic reticulum and endolysosomes, respectively [43-45].

Signaling pathways specific for CCK2R

CCK2R activates Epidermal growth factor receptor (EGFR) via PKC activated MMP3, which cleaves membrane attached pro-HBEGF into mature HBEGF [46, 47] . PKC isoforms PKC-β and PKC-η have been reported only downstream of CCK2R signaling [48, 49]. CCKR2 specific activation of β-catenin and E-cadherin is mediated by p21 protein-activated kinase 1 (PAK1) [50, 51] and CCK2R specific modulation of BCL-protein family signaling regulates mitochondrial cytochrome C release [52, 53]. CCK2R is reported to activate MAPK7 [54], an upstream regulator of transcription factors MEF-B,-C and D, and the PKC-η target PRKD2 [48], which enhances nuclear export of HDAC7 thereby relieving transcriptional repression of target genes like NR4A1 [55].

Global analysis of the CCKR map

The CCKR signaling pathways constitute a complex network comprising over 500 species and about 400 reactions. We sought to pinpoint key regulators in these pathways by

identifying signaling components (nodes) that display a high number of interactions with other network components as assessed by their node degree (i.e., the number of other nodes connected directly to it) (Materials and Methods and Figure S1). The global analysis of the CCKR map indicates scale free characteristics, with four protein kinases AKT1, SRC, PKC and PAK1 and the small GTPase HRAS among the 'hub' proteins and likely to be the central regulators of multiple signaling cascades (Table 2). AKT1, SRC, PKC, and PAK1 also rank among the top 6 node degrees in the PPI network constructed from direct physical interactions between CCKR model proteins (Table S1 in Additional file 2). The signaling reactions encoded in the literature-based CCKR model are thus paralleled by experimentally observed PPIs.

Modular representation of the CCKR signaling map

The CCKR model (Figure 1) was decomposed into 18 sub-network modules using the BiNoM tool (Figure 2A, Additional file 3, details of modules Rho GTPase and BCL in Figure 3). Each of these modules represents a structural and functional signaling subunit, combining a distinct set of closely coordinated molecular events concerning a particular protein or a protein complex. Details for all modules are in Additional file 4.

Table 2. Global analysis of the CCKR map

Rank	Species name	Species classification	Closeness centrality	Degree	Module assignment
1	AKT1	Kinase	0,144	12	AKT1
2	SRC	Kinase	0,133	12	SRC
3	HRAS	small GTPase	0,152	11	SRC
4	CCK2R	Receptor	0,142	11	CCK2R
5	PKC	Kinase	0,136	11	PKC
6	PAK1	Kinase	0,126	10	Rho GTPase

Involvement of different signaling modules in gastrin regulated cellular processes

Depending on cell types and the state of cells, gastrin can induce different cellular outcomes, such as proliferation, migration and apoptosis. While central modules such as PKC, AKT1, Rho GTPase, MAP3K11, MAPK1/3 and AP1 are reportedly involved in all three cellular outcomes, other signaling mechanisms are more specialized, e.g. the BCL-module signaling in apoptosis (Figure 2A).

Molecular mechanisms underlying gastrin-mediated proliferation involve regulation of protein synthesis and cell cycle. Protein translation is stimulated via the AKT1-module component mTOR triggering p70 S6 kinase [56, 57]. Gastrin-induced transcription of Cyclin D1, a central regulator of cell cycle progression, is mediated by JUN, FOS, CREB1, and TCF7L2 [58, 59], which are components of the modules: AP1, ATF2 and β-catenin. The modular representation (Figure 2A) depicts that EGFR-associated signaling enhances gastrin-induced proliferation by feed forward mechanisms involving SRC module components. Since the AKT1 module inhibits RAF1-MAPK1/3-module pathways by AKT1 kinase-mediated phosphorylation of RAF1, the involvement of AP1 and ATF2-module signaling in proliferation is more likely to proceed via SRC-MAP3K11.

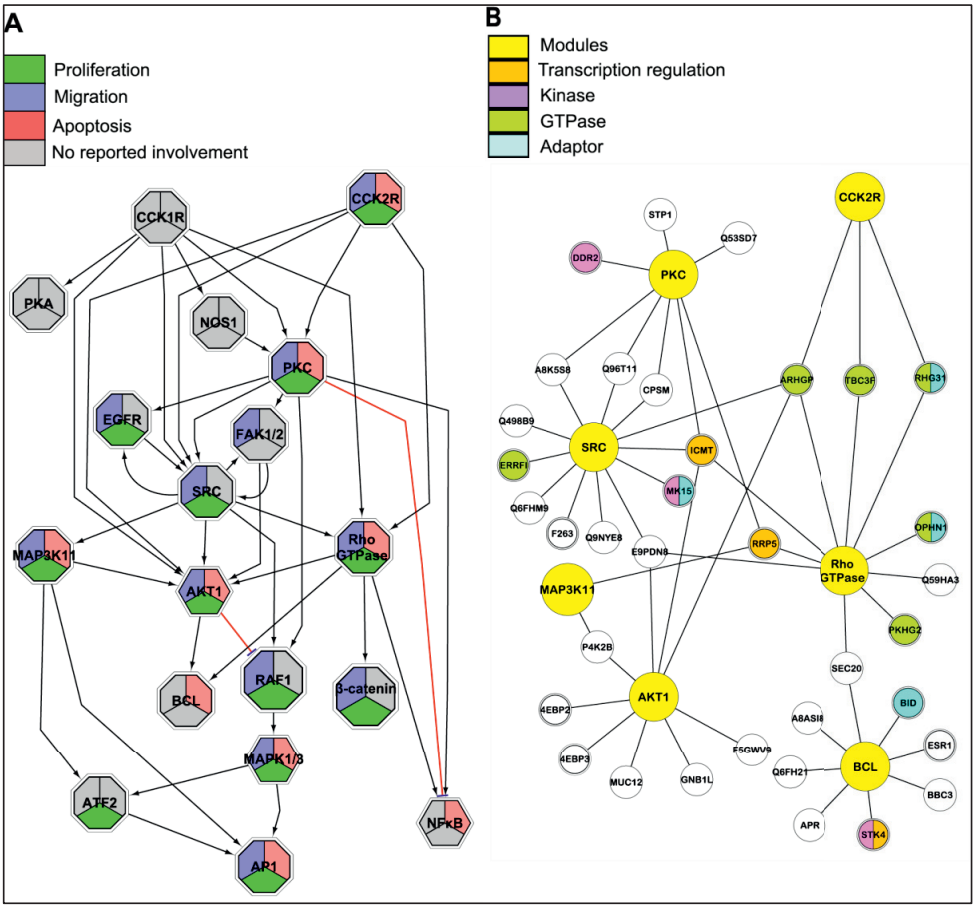


Figure 2. CCKR modular map and module-specific PathExpand Interactors. The modules are connected by 'activation' and 'inhibition' relationships derived from the detailed map thus representing central decision-making aspects. **A.** The modular representation comprises receptor-centered modules CCK1R, CCK2R, and EGFR, modules common to CCK1R and CCK2R (PKC, SRC, MAP3K11, MAPK1/3, RAF1, AKT1, NFκB, MAP3K11, Rho GTPase, FAK1/2) as well as the CCK1R-specific modules NOS1 and PKA; and CCK2R-specific modules BCL and β-catenin. Color coding depicts published experimentally documented information concerning involvement of the module encoded signalling mechanisms in gastrin-mediated regulation of cellular responses proliferation, migration and apoptosis. **B.** PathExpand interactors (Table S3 in Additional file 6) of the 15 modules that are not transcription factor centered. Names of these PathExpand interactors are given in Table 3.

Gastrin promotes migration by activating transcription of MMP7 and MMP9 [60, 61] via SNAI1, β-catenin, MAPK8 and JUN, in the β-catenin, MAP3K11, and AP1 modules respectively. Cell adhesion, tightly linked with cell migration, is regulated through components FAK1 and FAK2; Paxillin, Crk-associated substrate (CAS), and v-crk sarcoma virus CT10 oncogene (CRK) in the FAK1/2 module which is controlled via both PKC and SRC-modules, the latter exerting a positive feedback on the FAK1/2 module.

Table 3. List of PathExpand module interactors

UniProt_ID	GeneName	Gene symbol	Module assignment
ARHGP	Rho guanine nucleotide exchange factor (GEF) 25	ARHGEF25	CCK2R, Rho GTPase, SRC, AKT1
RHG31	Rho GTPase activating protein 31	ARHGAP31	CCK2R, Rho GTPase
TBC3F	TBC1 domain family, member 3F	TBC1D3F	CCK2R, Rho GTPase
ICMT	isoprenylcysteine carboxyl methyltransferase	ICMT	PKC, Rho GTPase, SRC, AKT1
Q96T11	cDNA FLJ14518, weakly similar to ANKYRIN R		PKC, SRC
CPSM	carbamoyl-phosphate synthase 1, mitochondrial	CPS1	PKC, SRC
A8K5S8	cDNA FLJ78047		PKC, SRC
RRP5	programmed cell death 11	PDCD11	PKC, Rho GTPase, MAP3K11
DDR2	discoidin domain receptor tyrosine kinase 2	DDR2	PKC
Q53SD7	Put uncharac RASGRP3 (RAS guanyl releasing protein 3	RASGRP3	PKC
STP1	transition protein 1	TNP1	PKC
E9PDN8	Guanine nucleotide exchange factor DBS	MCF2L	SRC, Rho GTPase, AKT1
MK15	mitogen-activated protein kinase 15	MAPK15	SRC
ERRFI	ERBB receptor feedback inhibitor 1	ERRFI1	SRC
Q498B9	ASXL1 protein	ASXL1	SRC
Q6FHM9	CD59 antigen, complement regulatory protein	CD59	SRC
Q9NYE8	Jak3 N-terminal-associated protein MAJN (Fragment)	MAJN	SRC
F263	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 3	PFKFB3	SRC
P4K2B	phosphatidylinositol 4-kinase type 2 beta	PI4K2B	AKT1, MAP3K11
F5GWV9	Mucin-12	MUC12	AKT1
MUC12	Mucin-12	MUC12	AKT1
4EBP2	eukaryotic translation initiation factor 4E binding protein 2	EIF4EBP2	AKT1
4EBP3	eukaryotic translation initiation factor 4E binding protein 3	EIF4EBP3	AKT1
GNB1L	guanine nucleotide binding protein (G protein), beta polypeptide 1-like	GNB1L	AKT1
SEC20	BCL2/adenovirus E1B 19kDa interacting protein 1	BNIP1	Rho GTPase, BCL
OPHN1	oligophrenin 1	OPHN1	Rho GTPase
PKHG2	pleckstrin homology domain cont, fam G (w RhoGef domain),2	PLEKHG2	Rho GTPase
Q59HA3	IQ motif containing GTPase activating protein 2 variant		Rho GTPase
ESR1	estrogen receptor 1	ESR1	BCL
BID	BH3 interacting domain death agonist	BID	BCL
STK4	serine/threonine kinase 4	STK4	BCL
Q6FH21			BCL
BBC3	BCL2 binding component 3	BBC3	BCL
A8ASI8	BH3 interacting domain death agonist	BID	BCL
APR	phorbol-12-myristate-13-acetate-induced protein 1	PMAIP1	BCL

Anti-apoptosis is induced by gastrin via several mechanisms including BCL-mediated repression of pro-apoptotic caspases and AP1-activated expression of Clusterin [52, 53, 62, 63]. The modular representation reveals that these cellular responses are regulated by both PKC independent and PKC dependent mechanisms. This applies to NFκB and its

downstream anti-apoptotic *BIRC2* and *BIRC3* target genes, which can be activated either directly by PKC or independently of PKC through the Rho GTPase module. Likewise, the AKT1-involvement in regulation of the BCL-module can be mediated by PKC dependent mechanisms or independently of PKC by the CCKR2 - Rho GTPase–pathway. Activation of AP1 on the other hand, seems to be strictly dependent on PKC which mediates its effect via either RAF1-MAPK1/3 or SRC-MAP3K11 cascades.

We note that the AKT1-module inhibits the RAF1-MAPK1/3-route to AP1-activation and in parallel enhances Rho GTPase activation of the BCL-module. Thus, AKT1 can potentially promote BCL-module apoptosis-regulating mechanisms and at the same time block MAPK1/3-mediated AP1-activation. In the latter configuration the cell will rely on MAP3K11 to bypass the inhibitory effect of AKT1 on AP1-mediated regulation of gene expression.

Extending the CCKR map with large-scale PPI data

The comprehensive CCKR signaling map (Figure 1) has been constructed with a knowledge-driven approach based on molecular reactions and interactions reported in the literature, thereby inevitably leaving significant gaps concerning signaling events and mechanisms that are as yet unstudied. The sparseness of the model is also reflected by the fact that ~90% of nodes have a degree ≤3 (Figure S1 in Additional file 2), while it is well known that signaling networks are generally more highly interconnected [64, 65]. We have therefore exploited large-scale PPI data to complement our signaling network scaffold. This data-driven strategy allowed us to access information that had not yet been related to gastrin or CCK responses.

We identified 4119 proteins with binary interactions with CCKR signaling proteins (Table S2 in Additional file 5). Among those 146 proteins are also CCKR model proteins. Of particular interest is a group of 74 proteins that satisfied the PathExpand topological criterion implying that each of them increase the compactness of the global CCKR signaling network [29], (Figure S2 in Additional file 2, Table S3 in Additional file 6). Interestingly, 42 proteins of this global PathExpand group are not known to participate in any pathway in the KEGG and/or Reactome databases (Table S2 in Additional file 5). A GO term overrepresentation analysis [66] showed that this global PathExpand group of interactors is enriched in molecular functions relating to protein kinases, protein phosphatases, and GTPase-regulators indicating that many of them can potentially regulate the CCKR pathway via phosphorylation-dephosphorylation mechanisms and by interfering with small GTPases signaling.

PathExpand analysis of each of the 18 modules with their protein interactors separately, identified 72 proteins (PathExpand module set) that are multiply linked with and thus increase compactness of individual modules; among which 33 candidates that were not predicted by the PathExpand analysis on the global CCKR model (Table S4 in Additional file 7). Figure 2B shows the PathExpand module interactors for the 15 intracellular CCKR signaling modules (i.e. not showing the modules centered around one of the transcription factors AP1, ATF2, NFkB). Eleven of these proteins are linked to more than one module and can contribute both to PKC-independent (e.g. ARGHGP, RHG31, TBC3F) and PKC dependent (e.g. ICMT) signaling routes. In contrast, 24 members of the PathExpand module interactor set are linked to only one module, suggesting they may act as preferential regulators of this module. For instance, the Mitogen activated kinase 15 (MK15) is a compactness increasing protein only for the SRC module, where it interacts with the two kinases SRC and CSK.

Use cases

CCKR map and genome-scale PPI generate hypotheses for refinement of Rho GTPase and BCL module mechanisms

In order to demonstrate potential use of the CCKR map and its PPI extensions, we here discuss putative novel signaling mechanisms involved in gastrin mediated activation of anti-apoptosis via Rho GTPase and BCL modules.

The activation of small GTPases, RHOA, RAC1, and CDC42 of the Rho GTPase module (Figure 3A) is promoted by guanine exchange factor proteins (GEFs), which trigger conversion of the inactive GDP-bound form of small GTPases to the active GTP-bound form, here exemplified by Leukemia-associated Rho guanine-nucleotide exchange factor (LARG) [67]. The inactive form of the small GTPases is restored by the GTPase-activating proteins (GAPs) that enhance hydrolyzation of the bound GTP, depicted here by deactivation of RHOA and RAC1 by GAPs Rho GTPase-activating protein 4 (ARHGAP4) and Regulator of G-protein signaling 2 (RGS2), respectively. Incoming components of this module are trimeric G-proteins, $G\alpha_q$ and $G\alpha_{13,}$ and outgoing components are 3 kinases: ROCK1, PAK1 AKT1.

The 724 interactors of Rho GTPase module components include 157 GTPase associated proteins, comprising more than 50 of each of GEF-type small GTPase family activators and GAP-type small GTPase family deactivators (Table 4). Several of these GTPase-regulating interactors increase compactness of Rho GTPase module signaling mechanisms according to the PathExpand method [29], including the GEF-type Oligophrenin 1 (OPHN1) and Pleckstrin homology domain containing, family G, member 2 (PKHG2), as well as GAP-type IQ motif containing GTPase activating protein (Q59HA3), Rho GTPase activating protein 31 (RHG31) and TBC1 domain family, member 3F (TBC3F) (see Figure 3B). All of these interactors also increase compactness of the global CCKR pathway.

Sixteen of the 121 kinase interactors bind to at least 2 different components of the Rho GTPase module, including p21 protein (Cdc42/Rac)-activated kinases 2 and -3 (PAK2, PAK3) as well as Mitogen-activated protein kinase kinase kinase 4 and -10 (M3K4, M3K10) (Figure 3B). These kinases seem to indicate a vast array of phosphorylation-events likely to be involved in co-ordinating the signaling proteins involved in this sub-module of the CCKR pathway. It is worth noting that a high proportion of the kinase interactors (e.g. INSR, PK3CA, and ABL2) are linked to the incoming components of the Rho GTPase module. Table 4 shows that half of all protein interactors bind to and thus potentially regulate one of the incoming components to modules. Furthermore, more than half of all kinase interactors of the CCKR map proteins are found in this 'incoming component' interactor group. We believe that this underscores the value of this subset of interactors for hypotheses concerning new regulatory components of the CCKR pathways.

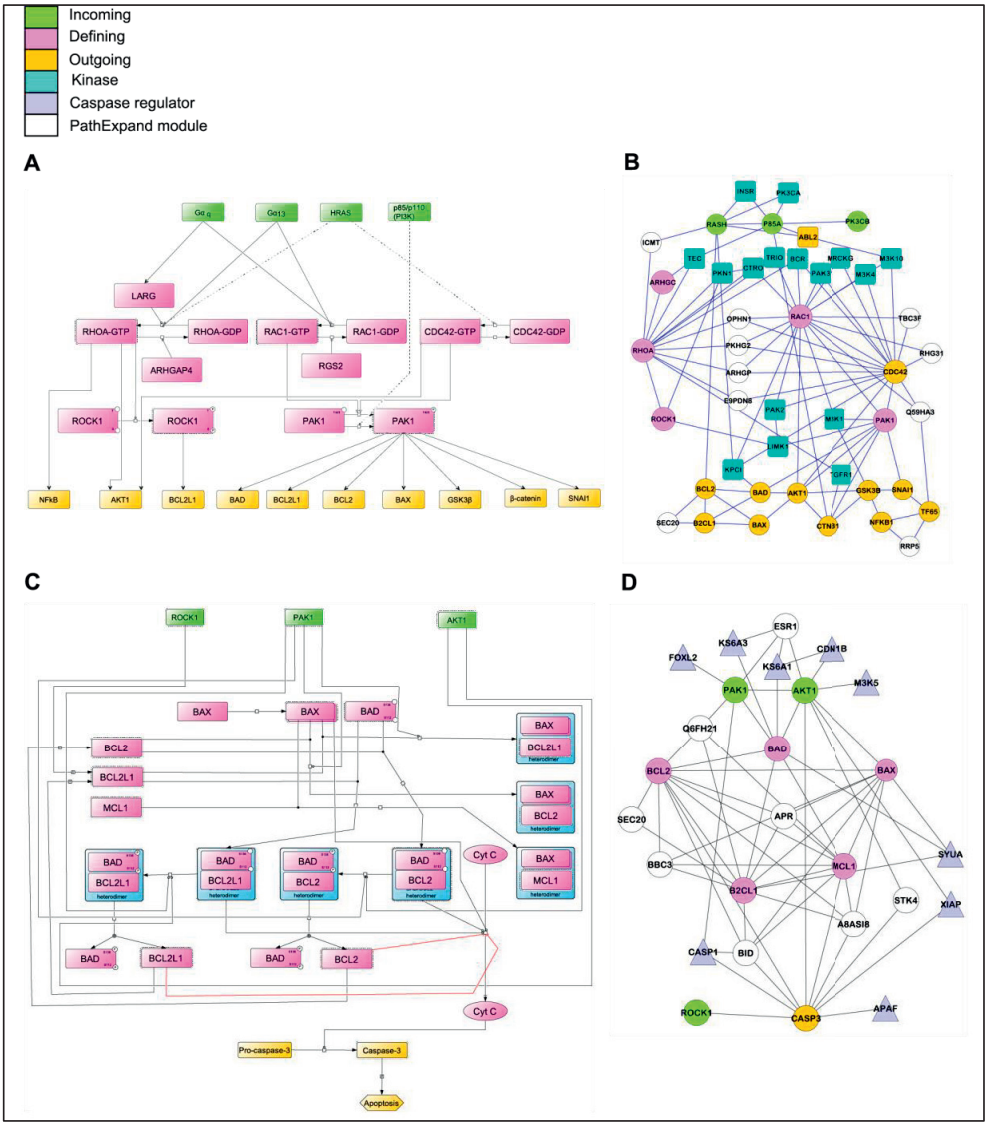


Figure 3. Rho GTPase and BCL modules and their interactors. Details of module **A.** Rho GTPase and **B.** BCL where 'defining' implies specific components within a module, 'incoming' as upstream regulators, and 'outgoing' as downstream effectors (Additional file 3). CCKR module component interactors for **C.** Rho GTPase and **D.** BCL modules are either colourless (PathExpand interactors) or colored according to GO molecular function classification (Table S5 in Additional file 8).

The BCL module controls release of cytochrome C from mitochondria, thereby regulating the activity of caspase 3 (Figure 3C). This signaling process is mediated by homo- and hetero-oligomerizations of the pro-apoptotic BCL family proteins BAX and BAD, and the anti-apoptotic BCL2, BCL2L1 and MCL1. Incoming regulators of the BCL

module are the gastrin-activated kinases ROCK1, PAK1 and AKT1 which mediate phosphorylation of BAD and BCL2L1 resulting in dissociation of BAD-BCL2 and BAD-BCL2L1 heterodimers, repression of cytochrome C release and, as a consequence inhibition of caspase 3 activity [53].

Among the PathExpand interactors of the BCL module (Figure 3D) we find a number of known BCL-family protein binding partners including BH3 interacting domain death agonist (BID), BCL2 binding component 3 (BBC3) and BCL2/adenovirus E1B 19kDa interacting protein 1 (SEC20) and Phorbol-12-myristate-13-acetate-induced protein 1 (APR). Serine/threonine kinase 4 (STK4), on the other hand, interacts with and is cleaved by BCL-module effector caspase 3 (CASP3) after which its cleavage products contribute both to enhancing downstream apoptosis and to inhibition of BCL module incoming component AKT1 [68]. Thus, PathExpand interactor STK4 can contribute to fine tuning of the gastrin regulated anti-apoptotic response.

BCL module interactors annotated with a GO caspase regulator term include Apoptotic peptidase activating factor 1 (APAF1), X-linked inhibitor of apoptosis (XIAP) and caspase 1 (CASP1) (Figure 3D). Physical interaction of CASP1 with CCKR model proteins PAK1 and CASP3 as well as with PathExpand interactor BID strongly suggests that caspase 1 may also be targeted in gastrin-mediated anti-apoptotic responses even though this has not yet been reported.

CCKR map extended with large-scale PPI used to generate hypotheses for dynamic gene regulatory networks

We next focus on transcriptional aspects of the CCKR pathway and show how kinetic models of transcription factor activity together with the CCKR map and its PPI extensions enables refinement of knowledge pertaining to molecular mechanisms involved in transcription regulation.

Our analyses of genome-scale temporal gene expression responses to a 14 hour gastrin treatment period in quiescent pancreatic adenocarcinoma AR42J cells, have identified temporal mRNA profiles of more than 2000 gastrin-responsive genes (Array Express accession number: GSE32869). Using Network Component Analysis (NCA) [69] we exploited these gastrin mRNA time series data to estimate the temporal transcription factor activity profiles of ~50 different transcription factors expressed in AR42J (unpublished data). Ten of these transcription factors (Figure 4A) are present in the CCKR-map, meaning that they reportedly are involved in gastrin- or CCK- mediated regulation of transcription. The NCA derived transcription factor activity profiles, which we have found to be of high accuracy (manuscript in preparation), indicate that transcription factors ATF2, AP1, EGR1, ELK1, NFκB, SRF, FOXO1 and FOXO3 are activated by gastrin already after 30-60 minutes (Figure 4B), while CREB1 activity displays a delayed peak at 2-4 hours (Figure 4C). TCF7L2 (also called TCF4) activity starts to increase after 4 hours and peaks at 10-12 hours. Furthermore, NFκB is estimated by NCA to exhibit a second activity peak at 10-12 hours (Figure 4D). The CCKR map indicates that upstream signaling mechanisms common to immediate early activated TFs involve several PKC-activated MAP kinases as well as AKT1. Together, these can explain the early activation of ATF2, AP1, EGR1, ELK1, and FOXO1 and 3 (Figure 4A). The early activation peak of NFκB is also likely to be mediated by PKC since we can infer from the other PKC-dependent responses, including rapid MAPK1/3 activation [70, 71] that PKC activation is an early event in the signaling process.

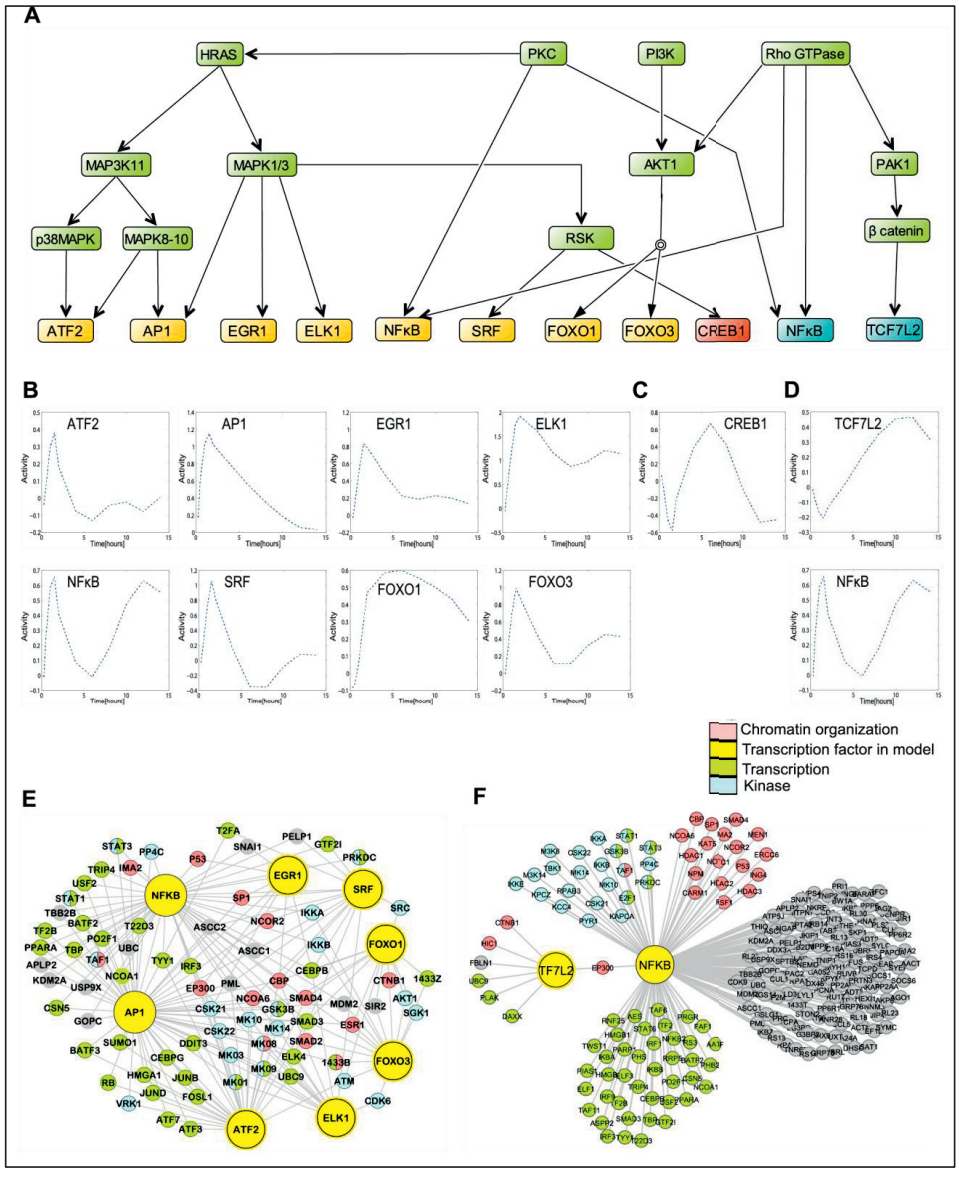


Figure 4. CCKR model transcription factor regulatory networks and activity profiles. The two dimer transcription factors, AP1 and NFκB, are constituted of JUN-FOS and NFκB1-RELA heterodimers, respectively. **A.** Upstream regulatory network (green) of CCKR model transcription factors (early: yellow, delayed: red, and late: blue). **B.** transcription factor activity temporal profile estimated by NCA: immediate early; **C.** delayed and; **D.** late active TFs. **E.** and **F.** Protein-protein interaction networks of AR42J expressed protein interactors with 'immediate early' and 'late' TFs, respectively.

Table 4. Overview of protein-protein interactors of the CCKR model

Interactor subsets	# interactors	Kinase[a]	Phosphatase[b]	GTPase-associated	Adaptor	Transcription regulation
All	4119	400	109	304	378	566
PathExpand_Global	74	6	9	8	7	3
PathExpand_Mod	72	7	7	6	3	21
PathExpand_all	106	10	11	8	7	22
Incoming_Mod	1974	276	71	206	233	239
AKT1	1154	198	51	148	165	145
AP1	643	84	45	37	61	163
ATF2	445	65	22	22	35	137
BCL	505	107	21	44	53	76
Beta-catenin	524	70	29	38	45	93
CCK1R	33	7	2	6	4	1
CCK2R	276	43	6	35	30	25
EGFR	567	88	22	45	103	57
FAK 1/2	966	139	34	81	151	100
MAP3K11	398	81	20	39	38	76
MAPK1/3	506	97	34	50	48	93
NOS1	162	45	9	7	33	10
NFκB	1041	116	23	91	88	156
PKA	239	52	11	15	29	31
PKC	566	116	24	55	88	74
RAF1	565	123	21	77	55	69
Rho GTPase	724	121	20	157	115	55
SRC	1097	163	39	98	166	116

[a]This includes kinase regulators
[b]This includes phosphatase regulators

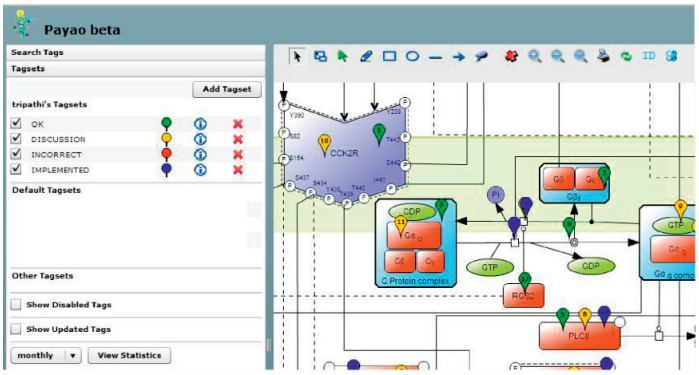


Figure 5. CCKR model curation in Payao. Part of the CCKR map Payao implicated with tagsets 'OK', 'DISCUSSION', 'INCORRECT,' and 'IMPLEMENTED' to record input from different curators on each reaction and components.

Interestingly, our CCKR map PPI extension (Figure 4E) confirms that the MAP kinases reported in literature to be involved in immediate early transcription factor activation (Figure 4A) are also engaged in direct physical interactions with these transcription factors. The PPI extension suggests that Casein kinase 2, alpha 1 and 2 (CSK21, CSK22) may also be involved in common mechanisms regulating the immediate early TFs ATF2, EGR1, FOS and JUN. Moreover, Figure 4E depicts extensive PPI among CCKR map transcription factors themselves as well as with additional transcription factors, transcription co-factors and chromatin modifiers, thus indicating the importance of devising goal directed approaches to experimentally address the functional interactome of transcription regulation.

Since our CCKR signaling pathway model has RSK as the only upstream regulator of both the immediate early TF SRF and the delayed CREB1, the available gastrin- and CCK-related literature is clearly insufficient to explain the differences in kinetic protein activity profiles of these two transcription factors. Searching the candidates for upstream signaling components we noted that CREB1 but not SRF interacts with the CCKR model protein Glycogen synthase kinase-3 beta (GSK3B), reported to inhibit CREB1 activation and DNA binding by phosphorylating other CREB amino acid residues than S133 [72, 73]. We speculate that if GSK3B is active only during the first hour of the gastrin response, this kinase can cause the delayed activation of CREB1.

Likewise, delayed activation of CREB1 has been observed to be associated with p38MAPK mediated activation of Ribosomal protein S6 kinase alpha-4 (KS6A4) and Ribosomal protein S6 kinase alpha-5 (KS6A5) [74, 75], and both KS6A4 and KS6A5 are found to be one of the interactors of CREB1 in our large scale PPI analysis. Further research on their role in gastrin mediated delayed activation of CREB1 is necessary to corroborate this.

Similarly, the difference in the temporal profiles of FOXO1, which is suggested to stay at high activity until at least 10h of the gastrin response, and FOXO3, whose protein activity returns to baseline within 6h, cannot be explained by our current model, since the only upstream regulator, AKT1 is common to both of them (Figure 4A). From our PPI CCKR model extensions we observe three potential upstream regulators, p21 protein (CDC42/RAC)-activated kinase 1 (PAK1), v-src (SRC) and Dual-specificity tyrosine-(Y)-phosphorylation regulated kinase 1A (DYRK1A) as well as four transcription regulators, CCAAT/enhancer binding protein (C/EBP) (CEBPB), CREB binding protein (CREBBP), Hepatocyte nuclear factor 4, alpha (HNF4A) and Peroxisome proliferator-activated receptor gamma (PPARG), which interact with FOXO1 but not with FOXO3 and that may be interesting to investigate for their potential involvement in the prolonged FOXO1 activation.

We hypothesized that the intracellular signaling mechanism responsible for the late NFκB activity peak could be Rho GTPase since the other late transcription factor, TCF7L2, is also downstream of Rho GTPase (Figure 4A). However, gastrin-mediated activation of Rho GTPase in AR42J cells is reported to occur very early, within 15 minutes [76]. Thus, our current CCKR map representation of the intracellular signaling mechanisms leading to NFκB-activation is not sufficient to explain the late phase of its biphasic activity profile. NFκB interacts with a large array of kinases, transcription regulators and chromatin modifiers (Figure 4F) that may potentially be involved in its late phase activation. One of several NFκB

partners that are transcriptionally upregulated by gastrin in AR42J cells, is Nuclear receptor coactivator 1 (NCOA1), with an mRNA peak at 6 hours. NCOA1 may thus enable NFκB mediated expression of the late gastrin-responding genes analogous to its observed potentiation of NFκB mediated regulation of other target genes [77, 78]. Likewise, upregulation of another NFκB-interacting protein CCAAT/enhancer-binding protein beta (CEBPB) [79, 80], which exhibits an extended peak spanning 2-8 hours of the gastrin response, may enable late NFκB-mediated gene expression.

For the late protein activity of TCF7L2, the PPI extension allows us to postulate that late activation of TCF7L2 partner Death-domain associated protein (DAXX) and DAZ-associated protein 2 (F8VU62) known to potentiate TCF7L2 transcription [81, 82], or late inactivation of TCF7L2 partner Hypermethylated in cancer 1 (HIC1) which represses TCF7L2 transcription [83] may play a role. Another mechanism that may be of interest to pursue in efforts to explain the TCF7L2 activity is the availability of its protein partner β-catenin (CTNB1). CTNB1, in its inactive state is sequestered and thus rendered inaccessible for TCF7L2 by amongst others E-cadherin (CADH1). In the early phase of gastrin response, E-cadherin is released from its transcriptional inhibition by transcription factor SNAI1. Thus, increased levels of E-cadherin during the first hours of gastrin response may decrease availability of CTNB1 for TCF7L2-mediated transcription.

Discussion

We present a map of signaling cascades mediated by two closely related receptors (CCK1R, CCK2R). We enhance the applicability of the map for hypothesis generation by two central strategies. First, we provide a computationally modularized version of topologically and functionally connected meta-nodes. This modular view simplifies navigation through the comprehensive CCKR map and provides for an improved, higher level comprehension of pathway regulatory aspects involved in cell fate decisions related to proliferation, migration and apoptosis. Molecular mechanistic insight into these cellular responses is of high importance for improved understanding of normo- and pathophysiological processes such as gastrin/CCKR2-linked carcinogenesis [47, 52, 84, 85], as well as for cholecystokinin induced hypoplasia, cell regeneration and digestive enzyme secretion [86]. Secondly, we take advantage of public large scale binary PPI knowledge to predict new potential regulators of the CCKR signaling, including 106 interactors that significantly enhance the compactness of the CCKR network [29], through tight direct and indirect interactions with model proteins. For the remaining close to 4000 interactors that do not comply with these strict connectivity requirements, we demonstrate the use of GO molecular function and CCKR map module interaction information (Figure 3) to identify specific subsets of potentially high interest for a more detailed perception of intracellular gastrin- and CCK-responses. Moreover, we show that the PPI data can be used to partly explain gastrin-induced temporal transcription factor activity in dynamic gene regulatory networks derived from the CCKR model and genome-scale gene expression time series. Although further experimental validations are needed to confirm these new CCKR signaling mechanisms, they represent an important source of high quality hypotheses as a first step to develop a better comprehension of CCKR pathways functionality.

The advantage of our strategy compared to other recently published computational approaches for high-throughput hypothesis generation [23, 87, 88] is the complementing approaches with i) biological background knowledge encoded in the signaling map, including in the modules, manually curated

from literature reporting detailed experimental analyses of gastrin- and CCK-signaling and ii) large-scale PPI information downloaded from available databases of interactions, and filtered for binary physical interaction based on detection methods.

Conclusion

Our work demonstrates how the integration of a comprehensive model of complex biological networks with multiple dimensions of genome scale data can provide new knowledge on molecular mechanisms underlying dynamic cellular processes. The provided SBML-version of the CCKR map can serve as a starting point to generate quantitative mathematical models [89] for simulation and prediction of cellular outcomes in response to perturbations of the network. Further development of the resources presented here should be of high interest in translational research aimed at identifying new targets and biomarkers for improved treatment of gastrin- and/or cholecystokinin-related disease, such as cancer.

Materials and Methods

1. Construction of the CCKR map

Below is an overview of the CCKR pathway reconstruction procedure:

i) CellDesigner 4.2 is a structured diagram editor for drawing gene-regulatory and biochemical networks, and uses the standardized technologies; Systems Biology Graphical Notation (SBGN) process diagrams [90] and Systems Biology Mark-up Language (SBML) [91]. MIRIAM (Minimum Information Requested In the Annotation of Models) was followed to characterize each species in the comprehensive map [92].

ii) Knowledge encoded in the CCKR map was obtained from scientific publications and from pathway databases http://www.pathguide.org/ [93]. The following strategy was adopted in order to assemble a comprehensive corpus of scientific publications for generation of the model:

a) Review articles were searched in PubMed using general search terms with different combinations of cholecystokinin (CCK)/CCK1R, gastrin/CCK2R or searched by expert's name. Although these reviews provide important information they in general lack experimental validation to substantiate the interactions. Hence we examined all relevant original articles involving experimental evidences for the interactions quoted in reviews.

b) For a more exhaustive and updated literature collection, we used literature-mining tools LitInspector (http://www.litinspector.org/) [94] and iHOP (http://www.ihop-net.org/UniPub/iHOP/) [95].

c) Next, we performed a manual search for additional literature in PubMed using search terms such as Cholecystokinin (CCK)/CCK1R, Gastrin/CCK2R, or searched by author's name.

d) Last, we checked all the citations of already collected articles in ISI Web of Knowledge (Thomson Reuters Web of Knowledge ^{SM}).

iii) CellDesigner species and the reaction "notes" feature were used to record PMID and cell-type specific information for each reaction and interacting component in the CCKR map.

iv) Final curation and quality control was done in a collaborative effort involving 5 different research group members using of the community curation platform Payao (http://www.payaologue.org) [31], which

enabled efficient exchange of comments and tags. Consensus and critical comments from each annotator about precise representation of reactions, components, size and its cellular localization were discussed and implemented (Figure 5). After implementing inputs from curators within the group, the CCKR model was published as open source for the whole scientific community working with Payao [31]. Thereby we hope to receive comments and tags from the community of curators to keep increasing the quality of this CCKR signaling pathway map and keep it up to date with our increasing biological understanding.

2. Global analysis of CCKR pathway

We performed network topology study of the CCKR map using Cytoscape version 2.8 [96]. For the global analysis of the intracellular cascades, we removed connections downstream of the transcription factors in the comprehensive CCKR map. Resulting SBML file was then imported in the Cytoscape using BiNOM plugin [28]. BiNOM considers both 'reaction' and 'species' of a CellDesigner map as a node. Therefore, the CCKR network when imported in Cytoscape had 807 nodes (475 species and 332 reactions) with 963 edges. Next, we calculated the network statistics using 'Network Analysis' plugin [97] in Cytoscape assuming the network as undirected. Number of nodes connected directly to each node is called its degree and this data for all nodes in the network is known as the degree distribution of the network (Figure S1 in Additional file 2). The nodes with highest degree are called 'hubs'. Further, we calculated the closeness centrality which implies how fast information is transferred from one node to any other node in the network. It is the reciprocal of the average of shortest path lengths a node has with other nodes. Hence, higher the closeness centrality (between 0-1), shortest is the distance with other nodes and faster is the information flow.

3) BiNoM construction of modules

The BiNoM software was used to import information from CellDesigner to Cytoscape [28, 98] and build a modular view of the CCKR pathway. This higher level pathway representation is fully based on the underlying detailed map and helps to navigate through it.

We used the 'prune the graph' function of BiNoM to automatically separate the strongly connected components (SCC) of the network from the input and output species. The SCC were decomposed into smallest sub-networks with function 'extract material components'. Next, subnetworks with 50 percent or more overlapping nodes were merged into a single subnetwork. We then compared the merged network from all modules with our original model for the completeness. Redundant nodes were deleted; orphan nodes were added to relevant modules. The main network and the modules are available as Cytoscape session file (Additional file 3).

4) Protein-protein interaction networks (PPI)

PPI data were downloaded from PSICQUIC (all databases, version June 2012), and filtered for binary physical interactions based on PSI-MI controlled vocabulary method descriptions, following the procedure in [99] (Charles E. Chapple, personal communication). 4119 interactors were identified for CCKR signaling proteins which we were able to map to a specific UniProt accession number (Table S2 in Additional file 5).

The PathExpand method was applied on the CCKR model proteins interactors using the complete PPI network as a background [29].

5) Network component analysis (NCA)

Network component analysis is a computational method for approximating transcription factor activities (TFAs), by reconstructing target gene expression data by

matrix calculation of known TF connectivity with assumed TF activity until convergence [69].NCA decomposition can be represented as

$$[E] = [C][T] \quad (1)$$

Where, [E] is expression matrix, [C] represents connectivity matrix and [T]corresponds to transcription factor activity matrix. Based on the above formulation, the decomposition of [E] into [C] and [T] can be achieved by minimizing the following objective function:

$$\min \| ([E] - [C][T]) \| \quad (2)$$

s.t. C ϵ Z0

In order to guarantee uniqueness of the solution for the equation (2) up to a scaling factor, NCA criteria must be satisfied which includes: (a) The connectivity matrix [C] must have full-column rank. (b) When a node in the regulatory layer is removed along with all of the output nodes connected to it, the resulting network must be characterized by a connectivity matrix that still has full-column rank. (c) T matrix must have full row rank. We used gastrin treated (0-14h) temporal gene expression data (Array Express accession number: GSE32869) as expression matrix and TF-TG relation data from TFactS [100] as connectivity matrix.

Availability of supporting data

The data sets supporting the results of this article are included within the article and its additional files.

Conflict of interest

The authors declare that they have no conflict of interest.

Author contributions

ST conceived the idea of the CCKR map, participated in pathway construction, curation and modular decomposition, interpretation of PPI- and NCA-analyses and manuscript writing. **ÅF** participated in pathway construction and -curation and manuscript writing. **KC** participated in modular decomposition BiNoM analysis, carried out network topology study and helped in manuscript writing. **AB** provided protein-protein interaction network analysis data and manuscript writing. **JND** and **NSB** participated in generating NCA-analyses. **TB** participated in pathways curation and manuscript writing. **LT** raised funding, participated pathway curation and helped to draft the manuscript, **MK** raised funding, participated in interpretation of PPI- and NCA-analyses and helped to draft the manuscript. **AL** raised funding, participated in pathway curation, interpretation of PPI- and NCA-analyses and helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank Charles E. Chapple for sharing his protein-protein interaction network and Enrico Glaab for "a la carte" PathExpand analyses. This work was supported by The Norwegian Cancer Society, The Liaison Committee between the Central Norway Regional Health Authority (RHA), the Norwegian University of Science and Technology (NTNU) and Sør-Trøndelag University College (HisT). The microarray and bioinformatics services were provided by the Genomics Core Facility, Norwegian University of Science and Technology, and NMC - a national technology platform supported by the functional genomics program (FUGE) of the Research Council of Norway.

References

1. Rehfeld JF: **The New Biology of Gastrointestinal Hormones**. *Physiological reviews* 1998, **78**(4):1087-1108.
2. Watson SA, Grabowska AM, El-Zaatari M, Takhar A: **Gastrin - active participant or bystander in gastric carcinogenesis?** *Nature reviews Cancer* 2006, **6**(12):936-946.
3. Little TJ, Horowitz M, Feinle-Bisset C: **Role of cholecystokinin in appetite control and body weight regulation**. *Obesity Reviews* 2005, **6**(4):297-306.
4. Saluja aK, Saluja M, Printz H, Zavertnik a, Sengupta a, Steer ML: **Experimental pancreatitis is mediated by low-affinity cholecystokinin receptors that inhibit digestive enzyme secretion.** *Proceedings of the National Academy of Sciences of the United States of America* 1989, **86**:8968-8971.
5. Gukovsky I, Cheng JH, Nam KJ, Lee OT, Lugea A, Fischer L, Penninger JM, Pandol SJ, Gukovskaya AS: **Phosphatidylinositide 3-kinase gamma regulates key pathologic responses to cholecystokinin in pancreatic acinar cells.** *Gastroenterology* 2004, **126**:554-566.
6. Dabrowski A, Grady T, Logsdon CD, Williams Ja: **Jun kinases are rapidly activated by cholecystokinin in rat pancreas both in vitro and in vivo.** *The Journal of biological chemistry* 1996, **271**:5686-5690.
7. Gibbs J, Young RC, Smith GP: **Cholecystokinin decreases food intake in rats.** *Journal of comparative and physiological psychology* 1973, **84**:488-495.
8. Witkamp RF: **Current and future drug targets in weight management.** *Pharmaceutical research* 2011, **28**:1792-1818.
9. Varga G, Bálint A, Burghardt B, D'Amato M: **Involvement of endogenous CCK and CCK1 receptors in colonic motor function.** *British journal of pharmacology* 2004, **141**:1275-1284.
10. Dufresne M, Seva C, Fourmy D: **Cholecystokinin and gastrin receptors**. *Physiological reviews* 2006, **86**(3):805-847.
11. Cawston EE, Miller LJ: **Therapeutic potential for novel drugs targeting the type 1 cholecystokinin receptor**. *British journal of pharmacology* 2010, **159**(5):1009-1021.
12. Konturek PC, Konturek SJ, Brzozowski T: **Helicobacter pylori infection in gastric cancerogenesis**. *Journal of physiology and pharmacology : an official journal of the Polish Physiological Society* 2009, **60**(3):3-21.
13. Matysiak-Budnik T, Mégraud F: **Helicobacter pylori infection and gastric cancer**. *European Journal of Cancer* 2006, **42**(6):708-716.
14. Noble F, Wank SA, Crawley JN, Bradwejn J, Seroogy KB, Hamon M, Roques BP: **International Union of Pharmacology. XXI. Structure, Distribution, and Functions of Cholecystokinin Receptors**. *Pharmacological Reviews* 1999, **51**(4):745-781.
15. Christophe J: **Pancreatic tumoral cell line AR42J: an amphicrine model**. *American Journal of Physiology - Gastrointestinal and Liver Physiology* 1994, **266**(6):G963-G971.
16. Ochsner SA, Watkins CM, LaGrone BS, Steffen DL, McKenna NJ: **Research Resource: Tissue-Specific Transcriptomics and Cistromics of Nuclear Receptor Signaling: A Web Research Resource**. *Molecular Endocrinology* 2010, **24**(10):2065-2069.
17. Diehl CJ, Barish GD, Downes M, Chou MY, Heinz S, Glass CK, Evans RM, Witztum JL: **Research Resource: Comparative Nuclear Receptor Atlas: Basal and Activated Peritoneal B-1 and B-2 Cells**. *Molecular Endocrinology* 2011, **25**(3):529-545.
18. Lee MJ, Ye AS, Gardino AK, Heijink AM, Sorger PK, MacBeath G, Yaffe MB: **Sequential Application of Anticancer Drugs Enhances Cell Death by Rewiring Apoptotic Signaling Networks**. *Cell* 2012, **149**(4):780-794.
19. Kaizu K, Ghosh S, Matsuoka Y, Moriya H, Shimizu-Yoshida Y, Kitano H: **A comprehensive molecular interaction map of the budding yeast cell cycle**. *Mol Syst Biol* 2010, **6**.
20. Gloaguen P, Crépieux P, Heitzler D, Poupon A, Reiter E: **Mapping the follicle-stimulating hormone-induced signalling networks**. *Frontiers in Endocrinology* 2011, **2**.
21. Oda K, Matsuoka Y, Funahashi A, Kitano H: **A comprehensive pathway map of epidermal growth factor receptor signaling**. *Mol Syst Biol* 2005, **1**:2005 0010.
22. Calzone L, Gelay A, Zinovyev A, Radvanyi F, Barillot E: **A comprehensive modular map of molecular interactions in RB/E2F pathway**. *Mol Syst Biol* 2008, **4**.
23. Caron E, Ghosh S, Matsuoka Y, Ashton-Beaucage D, Therrien M, Lemieux S, Perreault C, Roux PP, Kitano H: **A comprehensive map of the mTOR signaling network**. *Mol Syst Biol* 2010, **6**.
24. Fink MY, Pincas H, Choi SG, Nudelman G, Sealfon SC: **Research Resource: Gonadotropin-Releasing Hormone Receptor-Mediated Signaling Network in L beta T2 Cells: A Pathway-Based Web-Accessible Knowledgebase**. *Molecular Endocrinology* 2010, **24**(9):1863-1871.
25. Mizuno S, Iijima R, Ogishima S, Kikuchi M, Matsuoka Y, Ghosh S, Miyamoto T, Miyashita A, Kuwano R, Tanaka H: **AlzPathway: a comprehensive map of signaling pathways of Alzheimer's disease**. *BMC systems biology* 2012, **6**:52.
26. Patil S, Pincas H, Seto J, Nudelman G, Nudelman I, Sealfon SC: **Signaling network of dendritic cells in response to pathogens: a community-input supported knowledgebase**. *BMC systems biology* 2010, **4**:137.

27. Raza S, McDerment N, Lacaze PA, Robertson K, Watterson S, Chen Y, Chisholm M, Eleftheriadis G, Monk S, O'Sullivan M *et al*: **Construction of a large scale integrated map of macrophage pathogen recognition and effector systems**. *BMC systems biology* 2010, **4**:63.
28. Zinovyev A, Viara E, Calzone L, Barillot E: **BiNoM: a Cytoscape plugin for manipulating and analyzing biological networks**. *Bioinformatics* 2008, **24**(6):876-877.
29. Glaab E, Baudot A, Krasnogor N, Valencia A: **Extending pathways and processes using molecular interaction networks to analyse cancer genome data**. *BMC bioinformatics* 2010, **11**:597.
30. Funahashi A, Morohashi M, Kitano H, Tanimura N: **CellDesigner: a process diagram editor for gene-regulatory and biochemical networks**. *BIOSILICO* 2003, **1**(5):159-162.
31. Matsuoka Y, Ghosh S, Kikuchi N, Kitano H: **Payao: a community platform for SBML pathway model curation**. *Bioinformatics* 2010, **26**(10):1381-1383.
32. Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, Caudy M, Garapati P, Gopinath G, Jassal B *et al*: **Reactome: a database of reactions, pathways and biological processes**. *Nucleic Acids Res* 2011, **39**(Database issue):D691-697.
33. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M: **KEGG for integration and interpretation of large-scale molecular data sets**. *Nucleic Acids Research* 2012, **40**(D1):D109-D114.
34. Dufresne M, Seva C, Fourmy D: **Cholecystokinin and gastrin receptors.** *Physiological reviews* 2006, **86**:805-847.
35. Paillasse MR, de Medina P, Amouroux G, Mhamdi L, Poirot M, Silvente-Poirot S: **Signaling through cholesterol esterification: a new pathway for the cholecystokinin 2 receptor involved in cell growth and invasion**. *Journal of Lipid Research* 2009, **50**(11):2203-2211.
36. Sancho V, Berna MJ, Thill M, Jensen RT: **PKCθ activation in pancreatic acinar cells by gastrointestinal hormones/neurotransmitters and growth factors is needed for stimulation of numerous important cellular signaling cascades**. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* 2011, **1813**(12):2145-2156.
37. Quattrone A, Dewaele B, Wozniak A, Bauters M, Vanspauwen V, Floris G, Schoffski P, Chibon F, Coindre JM, Sciot R *et al*: **Promoting role of cholecystokinin 2 receptor (CCK2R) in gastrointestinal stromal tumours pathogenesis**. *The Journal of pathology* 2012.
38. Wu V, Yang M, McRoberts Ja, Ren J, Seensalu R, Zeng N, Dagrag M, Birnbaumer M, Walsh JH: **First intracellular loop of the human cholecystokinin-A receptor is essential for cyclic AMP signaling in transfected HEK-293 cells.** *The Journal of biological chemistry* 1997, **272**:9037-9042.
39. Sabbatini ME, Bi Y, Ji B, Ernst Sa, Williams Ja: **CCK activates RhoA and Rac1 differentially through Galpha13 and Galphaq in mouse pancreatic acini.** *American journal of physiology Cell physiology* 2010, **298**:C592-601.
40. Le Page SL, Bi Y, Williams Ja: **CCK-A receptor activates RhoA through G alpha 12/13 in NIH3T3 cells.** *American journal of physiology Cell physiology* 2003, **285**:C1197-1206.
41. Moustafa A, Sakamoto KQ, Habara Y: **A fundamental role for NO-PLC signaling pathway in mediating intracellular Ca2+ oscillation in pancreatic acini.** *Nitric oxide : biology and chemistry / official journal of the Nitric Oxide Society* 2011, **24**:139-150.
42. Cordelier P, Estève JP, Rivard N, Marletta M, Vaysse N, Susini C, Buscail L: **The activation of neuronal NO synthase is mediated by G-protein betagamma subunit and the tyrosine phosphatase SHP-2.** *The FASEB journal : official publication of the Federation of American Societies for Experimental Biology* 1999, **13**:2037-2050.
43. Thorn P, Gerasimenko O, Petersen OH: **Cyclic ADP-ribose regulation of ryanodine receptors involved in agonist evoked cytosolic Ca2+ oscillations in pancreatic acinar cells.** *The EMBO journal* 1994, **13**:2038-2043.
44. Cosker F, Cheviron N, Yamasaki M, Menteyne A, Lund FE, Moutin M-J, Galione A, Cancela J-M: **The ecto-enzyme CD38 is a nicotinic acid adenine dinucleotide phosphate (NAADP) synthase that couples receptor activation to Ca2+ mobilization from lysosomes in pancreatic acinar cells.** *The Journal of biological chemistry* 2010, **285**:38251-38259.
45. Calcraft PJ, Ruas M, Pan Z, Cheng X, Arredouani A, Hao X, Tang J, Rietdorf K, Teboul L, Chuang K-t *et al*: **NAADP mobilizes calcium from acidic organelles through two-pore channels.** *Nature* 2009, **459**:596-600.
46. Sinclair NF, Ai W, Raychowdhury R, Bi M, Wang TC, Koh TJ, McLaughlin JT: **Gastrin regulates the heparin-binding epidermal-like growth factor promoter via a PKC/EGFR-dependent mechanism**. *American Journal of Physiology - Gastrointestinal and Liver Physiology* 2004, **286**(6):G992-G999.
47. Miyazaki Y, Shinomura Y, Tsutsui S, Zushi S, Higashimoto Y, Kanayama S, Higashiyama S, Taniguchi N, Matsuzawa Y: **Gastrin induces heparin-binding epidermal growth factor-like growth factor in rat gastric epithelial cells transfected with gastrin receptor**. *Gastroenterology* 1999, **116**(1):78-89.
48. Sturany S, Van Lint J, Gilchrist A, Vandenheede JR, Adler G, Seufferlein T: **Mechanism of Activation of Protein Kinase D2(PKD2) by the CCKB/Gastrin Receptor**. *Journal of Biological Chemistry* 2002, **277**(33):29431-29436.
49. Yassin RR, Little KM: **Early signalling mechanism in colonic epithelial cell response to gastrin**. *The Biochemical journal* 1995, **311 (Pt 3)**:945-950.
50. He H, Shulkes A, Baldwin GS: **PAK1 interacts with beta-catenin and is required for the regulation of the beta-catenin signalling pathway by gastrins**. *Biochimica et biophysica acta* 2008, **1783**(10):1943-1954.
51. Mishra P, Senthivinayagam S, Rana A, Rana B: **Glycogen Synthase Kinase-3beta regulates Snail and beta-catenin during gastrin-induced migration of gastric cancer cells**. *Journal of molecular signaling* 2010, **5**:9.

52. He H, Baldwin GS: **Rho GTPases and p21-activated kinase in the regulation of proliferation and apoptosis by gastrins**. *The international journal of biochemistry & cell biology* 2008, **40**(10):2018-2022.
53. He H, Yim M, Liu KH, Cody SC, Shulkes A, Baldwin GS: **Involvement of G proteins of the Rho family in the regulation of Bcl-2-like protein expression and caspase 3 activation by Gastrins**. *Cellular signalling* 2008, **20**(1):83-93.
54. Guo Y-S, Cheng J-Z, Jin G-F, Gutkind JS, Hellmich MR, Townsend CM: **Gastrin stimulates cyclooxygenase-2 expression in intestinal epithelial cells through multiple signaling pathways. Evidence for involvement of ERK5 kinase and transactivation of the epidermal growth factor receptor.** *The Journal of biological chemistry* 2002, **277**:48755-48763.
55. von Blume J, Knippschild U, Dequiedt F, Giamas G, Beck A, Auer A, Van Lint J, Adler G, Seufferlein T: **Phosphorylation at Ser244 by CK1 determines nuclear localization and substrate targeting of PKD2**. *The EMBO journal* 2007, **26**(22):4619-4633.
56. Seva C, Kowalski-Chauvel A, Daulhac L, Barthez C, Vaysse N, Pradayrol L: **Wortmannin-Sensitive Activation of p70S6-Kinase and MAP-Kinase by the G Protein-Coupled Receptor, G/CCKB**. *Biochemical and biophysical research communications* 1997, **238**(1):202-206.
57. Kikani CK, Dong LQ, Liu F: **"New"-clear functions of PDK1: beyond a master kinase in the cytosol?** *Journal of cellular biochemistry* 2005, **96**(6):1157-1162.
58. Pradeep A, Sharma C, Sathyanarayana P, Albanese C, Fleming JV, Wang TC, Wolfe MM, Baker KM, Pestell RG, Rana B: **Gastrin-mediated activation of cyclin D1 transcription involves beta-catenin and CREB pathways in gastric cancer cells**. *Oncogene* 2004, **23**(20):3689-3699.
59. Steigedal TS, Bruland T, Misund K, Thommesen L, Laegreid A: **Inducible cAMP early repressor suppresses gastrin-mediated activation of cyclin D1 and c-fos gene expression**. *American journal of physiology Gastrointestinal and liver physiology* 2007, **292**(4):G1062-1069.
60. Bierkamp C, Kowalski-Chauvel A, Dehez S, Fourmy D, Pradayrol L, Seva C: **Gastrin mediated cholecystokinin-2 receptor activation induces loss of cell adhesion and scattering in epithelial MDCK cells**. *Oncogene* 2002, **21**(50):7656-7670.
61. Mishra P, Senthivinayagam S, Rangasamy V, Sondarva G, Rana B: **Mixed Lineage Kinase-3/JNK1 Axis Promotes Migration of Human Gastric Cancer Cells following Gastrin Stimulation**. *Molecular Endocrinology* 2010, **24**(3):598-607.
62. Fjeldbo CS, Bakke I, Erlandsen SE, Holmseth J, Laegreid A, Sandvik AK, Thommesen L, Bruland T: **Gastrin upregulates the prosurvival factor secretory clusterin in adenocarcinoma cells and in oxyntic mucosa of hypergastrinemic rats**. *American journal of physiology Gastrointestinal and liver physiology* 2012, **302**(1):G21-33.
63. Ramamoorthy S, Stepan V, Todisco A: **Intracellular mechanisms mediating the anti-apoptotic action of gastrin**. *Biochemical and biophysical research communications* 2004, **323**(1):44-48.
64. Durmu, #351, Tekir S, #220, mit P, Eren Toku A, lgen K, #214: **Reconstruction of Protein-Protein Interaction Network of Insulin Signaling in Homo Sapiens**. *Journal of Biomedicine and Biotechnology* 2010, **2010**.
65. Paris L, Bazzoni G: **The protein interaction network of the epithelial junctional complex: a system-level analysis**. *Mol Biol Cell* 2008, **19**(12):5409-5421.
66. Maere S, Heymans K, Kuiper M: **BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks**. *Bioinformatics* 2005, **21**(16):3448-3449.
67. Booden MA, Siderovski DP, Der CJ: **Leukemia-associated Rho guanine nucleotide exchange factor promotes G alpha q-coupled activation of RhoA**. *Mol Cell Biol* 2002, **22**(12):4053-4061.
68. Cinar B, Fang PK, Lutchman M, Di Vizio D, Adam RM, Pavlova N, Rubin MA, Yelick PC, Freeman MR: **The pro-apoptotic kinase Mst1 and its caspase cleavage products are direct inhibitors of Akt1**. *Embo Journal* 2007, **26**(21):4523-4534.
69. Liao JC, Boscolo R, Yang YL, Tran LM, Sabatti C, Roychowdhury VP: **Network component analysis: reconstruction of regulatory signals in biological systems**. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**(26):15522-15527.
70. Hocker M: **Molecular mechanisms of gastrin-dependent gene regulation**. *Annals of the New York Academy of Sciences* 2004, **1014**:97-109.
71. Seufferlein T, Withers D, Broad S, Herget T, Walsh J, Rozengurt E: **The human CCKB/gastrin receptor transfected into rat1 fibroblasts mediates activation of MAP kinase, p74raf-1 kinase, and mitogenesis**. *Cell Growth Differ* 1995, **6**(4):383-393.
72. Hur EM, Zhou FQ: **GSK3 signalling in neural development**. *Nature reviews Neuroscience* 2010, **11**(8):539-551.
73. Grimes CA, Jope RS: **CREB DNA binding activity is inhibited by glycogen synthase kinase-3 beta and facilitated by lithium**. *J Neurochem* 2001, **78**(6):1219-1232.
74. Wu GY, Deisseroth K, Tsien RW: **Activity-dependent CREB phosphorylation: convergence of a fast, sensitive calmodulin kinase pathway and a slow, less sensitive mitogen-activated protein kinase pathway**. *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**(5):2808-2813.
75. Delghandi MP, Johannessen M, Moens U: **The cAMP signalling pathway activates CREB through PKA, p38 and MSK1 in NIH 3T3 cells**. *Cellular signalling* 2005, **17**(11):1343-1351.

76. Stepan V, Ramamoorthy S, Pausawasdi N, Logsdon CD, Askari FK, Todisco A: **Role of small GTP binding proteins in the growth-promoting and antiapoptotic actions of gastrin**. *American journal of physiology Gastrointestinal and liver physiology* 2004, **287**(3):G715-725.
77. Sheppard KA, Rose DW, Haque ZK, Kurokawa R, McInerney E, Westin S, Thanos D, Rosenfeld MG, Glass CK, Collins T: **Transcriptional activation by NF-kappaB requires multiple coactivators**. *Mol Cell Biol* 1999, **19**(9):6367-6378.
78. Na SY, Lee SK, Han SJ, Choi HS, Im SY, Lee JW: **Steroid receptor coactivator-1 interacts with the p50 subunit and coactivates nuclear factor kappaB-mediated transactivations**. *The Journal of biological chemistry* 1998, **273**(18):10831-10834.
79. El-Asmar B, Giner XC, Tremblay JJ: **Transcriptional cooperation between NF-kappaB p50 and CCAAT/enhancer binding protein beta regulates Nur77 transcription in Leydig cells**. *Journal of molecular endocrinology* 2009, **42**(2):131-138.
80. Dooher JE, Paz-Priel I, Houng S, Baldwin AS, Jr., Friedman AD: **C/EBPalpha, C/EBPalpha oncoproteins, or C/EBPbeta preferentially bind NF-kappaB p50 compared with p65, focusing therapeutic targeting on the C/EBP:p50 interaction**. *Molecular cancer research : MCR* 2011, **9**(10):1395-1405.
81. Huang YS, Shih HM: **Daxx positively modulates beta-catenin/TCF4-mediated transcriptional potential**. *Biochemical and biophysical research communications* 2009, **386**(4):762-768.
82. Lukas J, Mazna P, Valenta T, Doubravska L, Pospichalova V, Vojtechova M, Fafilek B, Ivanek R, Plachy J, Novak J *et al*: **Dazap2 modulates transcription driven by the Wnt effector TCF-4**. *Nucleic Acids Research* 2009, **37**(9):3007-3020.
83. Valenta T, Lukas J, Doubravska L, Fafilek B, Korinek V: **HIC1 attenuates Wnt signaling by recruitment of TCF-4 and beta-catenin to the nuclear bodies**. *Embo Journal* 2006, **25**(11):2326-2337.
84. Jun Cao J-PY, Chao-Hong Liu, Lan Zhou, Hong-Gang Yu: **Effects of gastrin 17 on β-catenin/Tcf-4 pathway in Colo320WT colon cancer cells**. *World journal of gastroenterology : WJG* 2006 **12**(46):7482-7487.
85. Clarke PA, Dickson JH, Harris JC, Grabowska A, Watson SA: **Gastrin enhances the angiogenic potential of endothelial cells via modulation of heparin-binding epidermal-like growth factor**. *Cancer research* 2006, **66**(7):3504-3512.
86. Trulsson LM, Gasslander T, Svanvik J: **Cholecystokinin-8-induced hypoplasia of the rat pancreas: influence of nitric oxide on cell proliferation and programmed cell death**. *Basic & clinical pharmacology & toxicology* 2004, **95**(4):183-190.
87. Leach SM, Tipney H, Feng W, Baumgartner WA, Kasliwal P, Schuyler RP, Williams T, Spritz RA, Hunter L: **Biomedical discovery acceleration, with applications to craniofacial development**. *PLoS Comput Biol* 2009, **5**(3):e1000215.
88. Gitter A, Carmi M, Barkai N, Bar-Joseph Z: **Linking the signaling cascades and dynamic regulatory networks controlling stress responses**. *Genome research* 2012.
89. Tiger CF, Krause F, Cedersund G, Palmer R, Klipp E, Hohmann S, Kitano H, Krantz M: **A framework for mapping, visualisation and automatic model creation of signal-transduction networks**. *Mol Syst Biol* 2012, **8**:578.
90. Novere NL, Hucka M, Mi H, Moodie S, Schreiber F, Sorokin A, Demir E, Wegner K, Aladjem MI, Wimalaratne SM *et al*: **The Systems Biology Graphical Notation**. *Nat Biotech* 2009, **27**(8):735-741.
91. Kitano H, Funahashi A, Matsuoka Y, Oda K: **Using process diagrams for the graphical representation of biological networks**. *Nature biotechnology* 2005, **23**(8):961-966.
92. Novere NL, Finney A, Hucka M, Bhalla US, Campagne F, Collado-Vides J, Crampin EJ, Halstead M, Klipp E, Mendes P *et al*: **Minimum information requested in the annotation of biochemical models (MIRIAM)**. *Nat Biotech* 2005, **23**(12):1509-1515.
93. Bader GD, Cary MP, Sander C: **Pathguide: a Pathway Resource List**. *Nucleic Acids Research*, **34**(suppl 1):D504-D506.
94. Frisch M, Klocke B, Haltmeier M, Frech K: **LitInspector: literature and signal transduction pathway mining in PubMed abstracts**. *Nucleic Acids Research* 2009, **37**(suppl 2):W135-W140.
95. Hoffmann R, Valencia A: **A gene network for navigating the literature**. *Nature genetics* 2004, **36**(7):664.
96. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T: **Cytoscape 2.8: new features for data integration and network visualization**. *Bioinformatics* 2011, **27**(3):431-432.
97. Assenov Y, Ramirez F, Schelhorn SE, Lengauer T, Albrecht M: **Computing topological parameters of biological networks**. *Bioinformatics* 2008, **24**(2):282-284.
98. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: A software environment for integrated models of biomolecular interaction networks**. *Genome research* 2003, **13**(11):2498-2504.
99. Souiai O, Becker E, Prieto C, Benkahla A, De Las Rivas J, Brun C: **Functional Integrative Levels in the Human Interactome Recapitulate Organ Organization**. *PLoS ONE* 2011, **6**(7).
100. Essaghir A, Toffalini F, Knoops L, Kallin A, van Helden J, Demoulin JB: **Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data**. *Nucleic Acids Research* 2010, **38**(11).

Additional files

Additional file 1 as XML document. The file is SBML version of the comprehensive CCKR map.

Additional file 2 as pdf. The file contains Figure S1, Figure S2, and Table S1. Figure S1 shows degree distribution of the CCKR map. Figure S2 illustrates PathExpand interactors of the CCKR map. Table S1 shows PPI based ranking of CCKR model proteins.

Additional file 3 as cys. This file is original cytoscape session file containing BiNoM generated modules of the CCKR map.

Additional file 4 as pdf. The file contains description of individual modules of the comprehensive CCKR map disentagled using BiNoM plugin in cytoscape.

Additional file 5 as xls. Table S2. The file contains list of CCKR model protein interactors identified from large scale protein-protein interaction network.

Additional file 6 as xls. Table S3. The file enlists PathExpand interactors of the global CCKR pathway.

Additional file 7 as xls. Table S4. The file depicts PathExpand analysis of BiNoM modules and their protein interactors.

Additional file 8 as xls. Table S5. The file encompasses list of selected Gene Ontology molecular function terms to classify large scale protein interactors.

Additional File 2

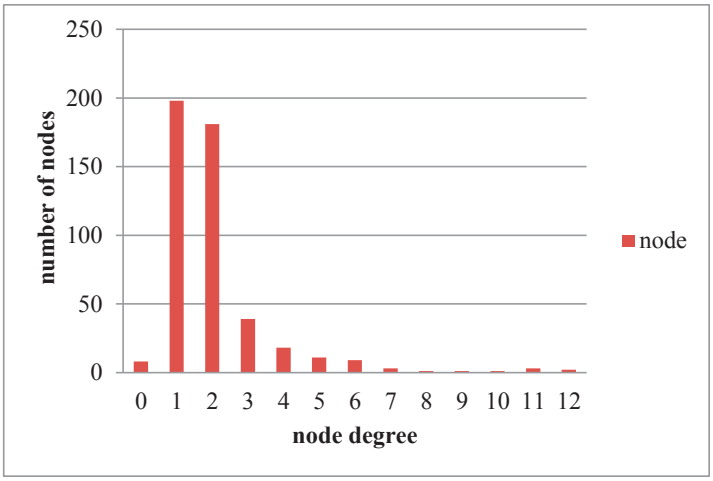


Figure S1. Degree distribution of the CCKR map

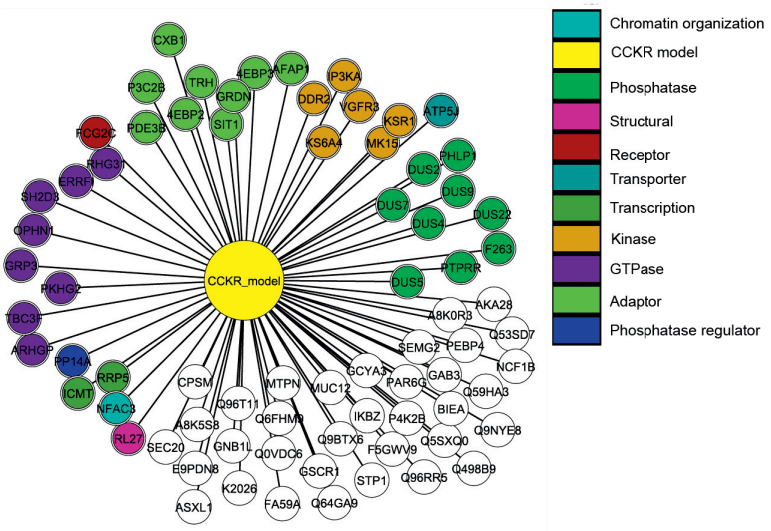


Figure S2. The protein kinase and phosphatase interactors that increase signaling pathway compactness are strong candidates to act as central regulators of the pathway via phosphorylation-dephosphorylation-based mechanisms and include kinases STK4, CSK21, CSK22, ITPKA, FLT4, DDR2, KS6KA4 and MK15 and a high number of Dual specificity phosphatases (DUSP1, 2, 4, 5, 7, 9, 22) in addition to PHLPP1 and PTPRR. The PECompact adaptors SH2D3 and (SIT1 are of high interest due to their potential to assist hypotheses on how CCKR model proteins are integrated with other proteins in the molecular (signaling) machinery. Of the 74 PECompact Global Interactors, 33 are not found to be PECompact for single CCKR pathway modules. See Table 3 of main text for PECompact Interactor names.

Table S1. PPI based ranking of the CCKR model proteins

Protein_model	**PPI**	**Rank**
GRB2_HUMAN	29	1
AKT1_HUMAN	26	2
MK01_HUMAN	26	2
MK08_HUMAN	26	2
SRC_HUMAN	26	2
RAF1_HUMAN	22	3
IKKB_HUMAN	19	4
IRS1_HUMAN	19	4
IKKA_HUMAN	18	5
KPCZ_HUMAN	18	5
P85A_HUMAN	18	5
KAPCA_HUMAN	17	6
KPCD_HUMAN	17	6
PAK1_HUMAN	17	6

Additional File 4

Description of BiNoM segmented modules

AKT1 module

AKT1 module encompasses components and reaction involved in activation of AKT1-mTOR cascade downstream of the CCKR (Table 1). Both gastrin and CCK activate PI3K via SRC dependent mechanism [101, 102]. Our model shows that gastrin activates tyrosine phosphorylation of IRS1 and its association with p85 subunit of PI3K by recruiting p85/p110 complex at the plasma membrane [103, 104]. Association of IRS1 with p85 activates PI3K complex. CCK2R stimulated JAK2 is documented to function upstream of PI3K in regulation of cell adhesion [105]. Active PI3K triggers PI3K dependent cascade by catalyzing PIP2 into PIP3. Activation of PI3K cascade promotes the recruitment of proteins with pleckstrin homology (PH) domains such as AKT1 and PDPK1 to the plasma membrane. Upon binding to the membrane, AKT1 and PDPK1 become active. Notably, translocation of AKT1 to the plasma membrane also facilitates its phosphorylation by PDPK1 [106, 107]. This cascade of events phosphorylates AKT1 at Ser308 and Ser473 to make it active. It is likely, that activated AKT1 regulates mTOR pathway and stimulates activation of Ribosomal protein S6 kinases 70kDa (p70 S6 kinase) because mTORC1 and PI3K specific inhibitor rapamycin, inhibits gastrin dependent p70S6K activity [56].

Table 1

Module_AKT1								
Incoming			Defining			Outgoing		
components	UniProt accession	UniProt KB entry	components	UniProt accession	UniProt KB entry	components	UniProt accession	UniProt KB entry
Gαq	P29992	GNA11_HUMAN	PRAS40	Q96B36	AKTS1_HUMAN	FOXO1	Q12778	FOXO1_HUMAN
SRC	P12931	SRC_HUMAN	IRS1	P35568	IRS1_HUMAN	FOXO3	O43524	FOXO3_HUMAN
FAK1	Q05397	FAK1_HUMAN	PDPK1	O15530	PDPK1_HUMAN	BAD	Q92934	BAD_HUMAN
CDC42	P60953	CDC42_HUMAN	p70S6K1	P23443	KS6B1_HUMAN	RPS6	P62753	RS6_HUMAN
RhoA	P61586	RHOA_HUMAN	mLST8	Q9BVC4	LST8_HUMAN	4E-BP1	Q13541	4EBP1_HUMAN
SHP2	Q06124	PTN11_HUMAN	mTOR	P42345	MTOR_HUMAN	EIF4E	P06730	IF4E_HUMAN
HRAS	P01112	RASH_HUMAN	PTEN	P60484	PTEN_HUMAN			
p38MAPK	Q16539	MK14_HUMAN	RHEB	Q15382	RHEB_HUMAN			
			RICTOR	Q6R327	RICTR_HUMAN			
			RPTOR	Q8N122	RPTOR_HUMAN			
			JAK2	O60674	JAK2_HUMAN			
			p85	P27986	P85A_HUMAN			
			p110	P42338	PK3CB_HUMAN			
			AKT1	P31749	AKT1_HUMAN			

AP1 module

AP1 module represents the life cycle of members of the AP1 transcription factor, JUN and FOS. Defining members of the AP1 module are listed in Table 2. Gastrin regulates JUN and FOS at both transcriptional and post-translational level. Gastrin mediated JUN gene transcription involves AP1 transcription factor whereas transcription of FOS gene is by MAPK dependent activation of ELK1 transcription factor [108]. At protein level, MAPK8 phosphorylates S63 and S73 residues of the JUN to make it active whereas ATF2 and MAPK1/3 activate FOS by phosphorylating its serine residues [108-110]. Active JUN and FOS protein translocate into the nucleus, associate together and form an AP1 complex.

Table 2

Module_AP1					
Incoming			Defining		
components	UniProt accession	UniProt KB entry	components	UniProt accession	UniProt KB entry
MAPK8	P45983	MK08_HUMAN	FOS	P01100	FOS_HUMAN
MAPK9	P45984	MK09_HUMAN	JUN	P05412	JUN_HUMAN
MAPK10	P53779	MK10_HUMAN			
p38MAPK	Q16539	MK14_HUMAN			
MAPK3	P27361	MK03_HUMAN			
MAPK1	P28482	MK01_HUMAN			

ATF2 module

This module explains the life cycle of ATF2 transcription factor. Members of the ATF2 module are listed in Table 3. Both MAPK8 and p38MAPK are associated with gastrin dependent activation of ATF2 by phosphorylating its threonine residues [108, 111, 112]. Active ATF2 forms a homodimer and translocates into the nucleus.Nuclear ATF2 associates with JUN to form a complex which regulates the transcription of JUN target gene [108].

Table 3

Module_ATF2					
Incoming			Defining		
components	UniProt accession	UniProt KB entry	components	UniProt accession	UniProt KB entry
MAPK8	P45983	MK08_HUMAN	ATF2	P15336	ATF2_HUMAN
MAPK9	P45984	MK09_HUMAN	JUN	P05412	JUN_HUMAN
MAPK10	P53779	MK10_HUMAN			
p38MAPK	Q16539	MK14_HUMAN			

BCL module

The BCL module includes BCL2 family proteins associated with gastrin dependent apoptosis regulation mechanism. In the comprehensive CCKR map, BCL2 family circumscribes both pro-apoptotic (BAX and BAD) and anti-apoptotic (BCL2, BCL2L1 and MCL1) members (Table 4). Gastrin activates expression of BCL2 and BCL2L1 via Rho GTPase dependent mechanism [53], whereas expression of MCL1 is mediated via AP1 dependent pathway [113]. Rho GTPase dependent activation of target proteins ROCK1 and PAK1 influences the expression of BCL2 and BCL2L1 proteins. Oligomerization of the BAX proteins causes release of cytochrome C from mitochondria, as a consequence activation of caspase 3. The BCL2-like proteins form heterodimers with BAX or BAD which results in the inhibition of the release of cytochrome C from the mitochondria [52, 53, 114]. Gastrin activates PAK1 and AKT1 proteins which then phosphorylate BAD at Ser136 and Ser112 residues, resulting in the dissociation of BAD from its heterodimer partners, BCL2 and BCL2L1 [53, 63, 115]. Dissociation of BAD stops release of cytochrome c from the mitochondria, as a consequence inhibits caspase 3 activation [53]. Gastrin inactivates both FOXO1 and FOXO3 transcription factors by AKT1 dependent phosphorylation [63] resulting in the inhibition of apoptosis.

Table 4

Module_BCL								
Incoming			Defining			Outgoing		
components	UniProt accession	UniProt KB entry	components	UniProt accession	UniProt KB entry	components	UniProt accession	UniProt KB entry
PAK1	Q13153	PAK1_HUMAN	BCL2L1	Q07817	B2CL1_HUMAN	Caspase 3	P42574	CASP3_HUMAN
ROCK1	Q13464	ROCK1_HUMAN	BAD	Q92934	BAD_HUMAN			
AKT1	P31749	AKT1_HUMAN	BAX	Q07812	BAX_HUMAN			
			BCL2	P10415	BCL2_HUMAN			
			MCL1	Q07820	MCL1_HUMAN			

Beta-catenin module

Beta-catenin module circumscribes components and reactions associated with gastrin dependent beta-catenin/E-cadherin interaction (Table 5). Beta catenin and E-cadherin form an adhesion complex at the membrane. This adhesion complex is disrupted by gastrin to promote cell migration and invasion. Gastrin activated PAK1 phosphorylates SNAI1 transcription factor at Ser246. Consequently, SNAI1 translocates to the nucleus and inhibits E-cadherin gene transcription [50, 51]. As a result, number of E-cadherin molecules present at the membrane decreases. Decrease in E-cadherin molecule causes disruption of Beta-catenin/E-cadherin interaction [50]. Now, intact Beta-catenin translocates into the cytoplasm and undergoes phosphorylation at Ser45 mainly by caseine kinase1 (CK1). Importantly, cytosolic phosphorylated beta-catenin may undergo GSK3beta dependent phosphorylation and eventual degradation but gastrin activated PAK1 prevent this degradation by inactivating GSK3beta. PAK1 inactivates GSK3beta by phosphorylating its Ser9 residue [50, 51]. Furthermore, PAK1 triggers nuclear transport of the activate beta-catenin [50, 51]. Nuclear beta-catenin associates with transcription factor TCF7L2 and regulate expression of several target genes in response to gastrin.

Table 5

Module_Beta-catenin								
Incoming			Defining			Outgoing		
components	UniProt accession	UniProt KB entry	components	UniProt accession	UniProt KB entry	components	UniProt accession	UniProt KB entry
PAK1	Q13153	PAK1_HUMAN	E-cadherin	P12830	CADH1_HUMAN	TCF7L2	Q9NQB0	TF7L2_HUMAN
			Beta-catenin	P35222	CTNB1_HUMAN			
			GSK3 beta	P49841	GSK3B_HUMAN			
			SNAI1	O95863	SNAI1_HUMAN			
			CK1	P48729	KC1A_HUMAN			

CCK1R module

The CCK1R module depicts the life cycle of the CCK1 receptor. Members of the CCK1R module are listed in Table 6. Sulfated CCK binds to active CCK1R and triggers downstream signaling cascades. Under CCK stimulation, CCK1R is rapidly phosphorylated at consensus serine residues in the third intracellular loop, both by PKC and a G protein kinase, causing receptor inactivation [116, 117]. Desensitization and further recycling of the CCK1R happens by receptor-endocytosis in the cytosol. After stimulation of CCK1R by CCK, ligand bound receptor complex is internalized into an endocytic vesicle [118, 119]. From the endosome, CCK and CCK1R are then sorted into their destined cellular location. Notably, average sorting time of CCK and receptor in endosome is about 25 minutes. CCK is sorted into the lysosome and undergoes proteosomal degradation whereas the receptor recycles back to the cell membrane with an average time of 60 min [119].

Table 6

Module_CCK1R								
Incoming			Defining			Outgoing		
components	UniProt accession	UniProt KB entry	components	UniProt accession	UniProt KB entry	components	UniProt accession	UniProt KB entry
CCK	P06307	CCKN_HUMAN	CCK1R	P32238	CCKAR_HUMAN	MAPK8	P45983	MK08_HUMAN
			Gαq	P29992	GNA11_HUMAN	MAPK9	P45984	MK09_HUMAN
			Gαs	P63092	GNAS2_HUMAN	MAPK10	P53779	MK10_HUMAN
			Gα13	Q14344	GNA13_HUMAN	p38MAPK	Q16539	MK14_HUMAN
						SHP2	Q06124	PTN11_HUMAN
						CD38	P28907	CD38_HUMAN

EGFR module

EGFR module encompasses components associated with the activation of EGFR (Table 7). Gastrin induces expression of HB-EGF and EGFR transactivation as documented in human gastric cancer cell line [120] and in rat gastric epithelial cells [46, 47]. CCK2R activates matrix metalloproteinase 3 (MMP-3) via PKC dependent mechanism [46]. Activated MMP-3 cleaves Glu-Asn site within the juxtamembrane (JM) region of the membrane anchored pro-HBEGF into soluble mature HBEGF [47, 120]. Mature HBEGF then binds to the EGFR and activates this receptor by phosphorylating several tyrosine residues. Both SRC and SHC1 associate with active EGFR. SHC1 binds to the tyrosine residues 1148/1173 of the active EGFR. EGFR activates SHC1 by phosphorylating its Y317 residue [121]. Also, GRB2 transports from cytosol and binds to the active EGFR receptor at membrane on either phosphorylated Y1068 or Y1086 [122]. This binding recruits SOS1 onto the membrane which forms a complex with GRB2. Phosphorylated SHC1 associates with GRB2/SOS1 complex and regulate CCK2R dependent MAPK cascade [123, 124].

Table 7

Module_EGFR								
Incoming			Defining			Outgoing		
components	UniProt accession	UniProt KB entry	components	UniProt accession	UniProt KB entry	components	UniProt accession	UniProt KB entry
SRC	P12931	SRC_HUMAN	EGFR	P00533	EGFR_HUMAN	GRB2	P62993	GRB2_HUMAN
			MMP3	P08254	MMP3_HUMAN	SOS1	Q07889	SOS1_HUMAN
			HBEGF	Q99075	HBEGF_HUMAN	SHC1	P29353	SHC1_HUMAN

CCK2R module

CCK2R represents the life cycle of CCK2 receptor. Members of the CCK2R module are listed in Table 8. Amidated gastrin binds to the CCK2R, which leads to the phosphorylation and activation of CCK2R. Active ligand-receptor complex triggers downstream signaling cascades. Internalization and intracellular trafficking of the CCK2R primarily involves binding of beta-arrestin adaptor proteins to the receptor and clathrin coated pits. After CCK2R stimulation by gastrin, beta-arrestin ½ transports from cytoplasm to the plasma membrane where it interacts with C-terminal phosphorylated residues of the CCK2 receptor [125]. Beta-arrestin bound CCK2R is then recuited into clathrin-coated endocytic vesicle. Interestingly, gastrin too found to be trapped into this endocytic vesicle but without any clear evidence whether it remains intact with the receptor or they were degraded by proteases [125]. It was examined that CCK2R internalization is also dependent on the activity of a GTPase, dynamin. Dynamin acts as a mechanochemical enzyme to clip membrane attached vesicles and their targeting, fusion with another compartment. Furthermore, internalized CCK2R does not recycle rapidly to the cell surface. Instead, CCK2R directs to the late endosome/lysosome, indicating a possibility of slow recycling/degradation.

Table 8

Module_CCK2R								
Incoming			Defining			Outgoing		
components	UniProt accession	UniProt KB entry	components	UniProt accession	UniProt KB entry	components	UniProt accession	UniProt KB entry
G17	P01350	GAST_HUMAN	Beta-arrestin-1	P49407	ARRB1_HUMAN	PLCγ1	P19174	PLCG1_HUMAN
CCK	P06307	CCKN_HUMAN	Beta-arrestin-2	P32121	ARRB2_HUMAN	PLCβ	Q9NQ66	PLCB1_HUMAN
			Dynamin	Q05193	DYN1_HUMAN	TRAF6	Q9Y4K3	TRAF6_HUMAN
			RGS2	P41220	RGS2_HUMAN	PLA2	P47712	PA24A_HUMAN
			CCK2R	P32239	GASR_HUMAN	MAP2K5	Q13163	MP2K5_HUMAN
			Gαq	P29992	GNA11_HUMAN	RAC1	P63000	RAC1_HUMAN
			AP2M1	Q96CW1	AP2M1_HUMAN	CDC42	P60953	CDC42_HUMAN
			PLA2	P47712	PA24A_HUMAN	JAK2	O60674	JAK2_HUMAN
						SHP2	Q06124	PTN11_HUMAN
						LARG	Q9NZN5	ARHGC_HUMAN
						SRC	P12931	SRC_HUMAN

FAK1/2 module

Tyrosine phosphorylation and activation of FAK1 is described in response to both CCK1R and CCK2R stimulation while FAK2 activation is recorded only in response to CCK. Gastrin controls cell adhesion by signaling pathways involving FAK1, Paxillin [126], Crk-associated substrate (CAS), and v-crk sarcoma virus CT10 oncogene (CRK) [126-128]. Gastrin activates CAS/CRK complex formation by p60SRC and PKC dependent pathway [128] (detail list in Table 9). Our model represents that FAK1 associates with p60SRC and p190RhoGEF to form a complex. This complex then regulates phosphorylation and activation of paxillin [126]. Interestingly, FAK1-p60SRC complex acts upstream of the gastrin-stimulated PI 3-kinase pathway [102]. In rat pancreatic acinar cells, CCK-8 rapidly stimulates tyrosine phosphorylation and activation FAK2. This activation of FAK2 is mediated by PKC and increase of [Ca^{2+}] [129, 130]. CCK stimulation causes a rapid formation of both FAK2-GRB2 and FAK2-CRK complexes [129]. The exact mechanism of FAK2 activation by Ca^{2+} is still not understood [131], but inhibiting PKC-θ in rat pancreatic acinar cells has been shown to inhibit phosphorylation of Tyr-402 of FAK2 [36], indicating that PKC-θ is the link between CCK1R stimulation and FAK2 activation. FAK2 auto-phosphorylates at Tyr402. Phosphorylation at Tyr402 provides a binding site for SH2 containing proteins including SRC and p85. Binding of SRC leads to phosphorylation of FAK2 residues Tyr579 and Tyr580, with maximal FAK2 kinase activity [132, 133]. Phosphorylation at Tyr-881 by SRC promotes interaction of FAK2 with GRB2 [133]. FAK2 also forms a complex with CRK in rat pancreatic acinar cells after stimulation with CCK-8 [129]. In rat pancreatic acinar cells CCK also stimulates formation of CRK-CAS complex [134].

Table 9

Module_FAK1/2								
Incoming			Defining			Outgoing		
components	UniProt accession	UniProt KB entry	components	UniProt accession	UniProt KB entry	components	UniProt accession	UniProt KB entry
SRC	P12931	SRC_HUMAN	CAS	P56945	BCAR1_HUMAN	HRAS	P01112	RASH_HUMAN
GRB2	P62993	GRB2_HUMAN	CRK	P46108	CRK_HUMAN			
PKC-θ	Q04759	KPCT_HUMAN	FAK1	Q05397	FAK1_HUMAN			
			FAK2	Q14289	FAK2_HUMAN			
			Paxillin	P49023	PAXI_HUMAN			
			p190RhoGEF	Q8N1W1	RGNEF_HUMAN			

MAPK1/3 module

MAPK1/3 module constitutes cascade of events associated with the activation of MAPK1 (ERK2) and MAPK3 (ERK1) signaling pathway downstream of CCKR (Table 10). CCKR activate MAPK1/3 through RAF dependent stimulation of MAP2K1/2.Activation of RAF is achieved either by GRB2/SOS dependent activation of HRAS or by stimulation of RAF through PKC-mediated mechanisms. Activation of RAF1 is independent of PKC activity in Rat1 cells whereas in human gastric cancer cells, RAS independent activation of RAF is detected in response to gastrin [70, 71].Active RAF1 phosphorylates serine residues of dual specificity kinases, MAP2K1 and MAP2K2. Phosphorylated MAP2K1 and MAP2K2 then activate MAPK1/3 by phosphorylating their threonine/tyrosine residues. Active MAPK1/3 then form a homodimer and transports into the nucleus where they regulate the activity of several TFs and TGs. Active MAPK1/3 also triggers RSK (RSK1/2) activation cascade by phosphorylating its serine/threonine residues [135]. Active RSK translocates into the nucleus where it plays a role in the activation of CREB1 TF by phosphorylating its S133 residue [135, 136]. Our modular view indicates that MAPK1/3 module has positive influence on AP1 and ATF2 modules.

Table 10

Module_MAPK1/3								
Incoming			Defining			Outgoing		
components	UniProt accession	UniProt KB entry	components	UniProt accession	UniProt KB entry	components	UniProt accession	UniProt KB entry
ARAF	P10398	ARAF_HUMAN	MAPK3	P27361	MK03_HUMAN	CREB1	P16220	CREB1_HUMAN
BRAF	P15056	BRAF_HUMAN	RSK1	Q15418	KS6A1_HUMAN	ELK1	P19419	ELK1_HUMAN
RAF1	P04049	RAF1_HUMAN	MAPK1	P28482	MK01_HUMAN	SP1	P08047	SP1_HUMAN
			MAP2K2	P36507	MP2K2_HUMAN	SP3	Q02447	SP3_HUMAN
			RSK2	P51812	KS6A3_HUMAN	SAP1	P28324	ELK4_HUMAN
			MAP2K1	Q02750	MP2K1_HUMAN	EGR1	P18146	EGR1_HUMAN
						SRF	P11831	SRF_HUMAN
						PPARγ	P37231	PPARG_HUMAN
						FOS	P01100	FOS_HUMAN

MAP3K11 (MLK3) module

In the modular view, MAP3K11 module represents components and reactions involved in the activation of MAPK8 (JNK) and p38MAPK (Table 11). Activation of MAPK8 and p38MAPK is reported in response to both CCK1R and CCK2R stimulation by CCK and gastrin respectively. MAP3K11 is a serine/threonine kinase with SH3 domain-containing proline-rich kinase. HRAS seems to be the upstream regulator

of MAP3K11. MAP3K11 is a known activator of dual specificity protein kinases MAP2K4 and MAP2K6 by phosphorylating Ser257/Thr261 residues of MAP2K4, and Ser207/Thr211 residues of MAP2K6. Phosphorylated MAP2K4 and MAP2K6 are the upstream regulators of MAPK8 and p38MAPK respectively [128, 137, 138]. MAP2K4 phosphorylates Thr183/Tyr185 residues of MAPK8 whereas MAP2K6 phosphorylates Thr180/Tyr182 residues of p38MAPK. Both MAPK8 and p38MAPK activate transcription factor ATF2 whereas only MAPK8 stimulate transcription factor JUN in RIE-1/CCK2R cells treated with gastrin [137].

Table 11

Module_MAP3K11								
Incoming			Defining			Outgoing		
components	UniProt accession	UniProt KB entry	components	UniProt accession	UniProt KB entry	components	UniProt accession	UniProt KB entry
HRAS	P01112	RASH_HUMAN	MAPK8	P45983	MK08_HUMAN	AKT1	P31749	AKT1_HUMAN
			MAP3K11	Q16584	M3K11_HUMAN	NFkB1	P19838	NFKB1_HUMAN
			p38MAPK	Q16539	MK14_HUMAN	RELA	Q04206	TF65_HUMAN
			MAP2K4	P45985	MP2K4_HUMAN	JUN	P05412	JUN_HUMAN
			MAP2K6	P52564	MP2K6_HUMAN	ATF2	P15336	ATF2_HUMAN
			MAPK9	P45984	MK09_HUMAN	MAPKAP-K2	P49137	MAPK2_HUMAN
			MAPK10	P53779	MK10_HUMAN	MEF2B	Q02080	MEF2B_HUMAN
			MAPKAP-K2	P49137	MAPK2_HUMAN	MEF2C	Q06413	MEF2C_HUMAN
						MEF2D	Q14814	MEF2D_HUMAN

NFκB module

This module encircles CCKR dependent activation mechanism of the NFκB transcription factor (Table 12). CCK1R and CCK2R dependent activation of NFκB is via PKCδ [108, 139]. It has been reported that both PKCδ and PKCε are involved in CCK mediated NFκB activation [139], and our model suggests that PRKD1 could be the possible link between this activation [140]. Gastrin stimulated CCK2R follows TRAF6/TAK1/MAP3K14 pathway to activate NFκB [108]. In this cascade, MAP3K14 which is also known as NFκB inducing kinase (NIK) activates IκB kinase. Activated IκB kinase then phosphorylates S32 and S36 residues of the IκB, as a result releases the inhibitory effect of IκB on the NFκB1-RELA complex. Phosphorylated IκB dissociates from the NFκB1-RELA complex and undergoes proteosomal degradation [108], leaving active NFκB1-RELA complex which then translocates into the nucleus. Gastrin and cholecystokinin promote NFκB nuclear translocation via RhoA and MAPK8 dependent pathways respectively [141, 142]. In CCK1R system, it has been observed that PKC-*α* exerts an inhibitory effect on NF*κ*B activation in rat pancreatic acini [139].

Table 12

Module_NFκB					
Incoming			Defining		
components	UniProt accession	UniProt KB entry	components	UniProt accession	UniProt KB entry
PKCα	P17252	KPCA_HUMAN	IKKα	O15111	IKKA_HUMAN
PKCδ	Q05655	KPCD_HUMAN	TRAF6	Q9Y4K3	TRAF6_HUMAN
PKCε	Q02156	KPCE_HUMAN	MAP3K14	Q99558	M3K14_HUMAN
RhoA	P61586	RHOA_HUMAN	IKKγ	Q9Y6K9	NEMO_HUMAN
PRKD1	Q15139	KPCD1_HUMAN	NFKB1	P19838	NFKB1_HUMAN
			TAK1	P49116	NR2C2_HUMAN
			IKKβ	O14920	IKKB_HUMAN
			RELA	Q04206	TF65_HUMAN

NOS1 module

The nitric oxide-cGMP pathway is known to be activated by CCK in rat pancreatic acinar cells [41] and in the CHO cell line by CCK1R [42, 143, 144]. Memebers of the NOS1 module are listed in Table 13. The link between CCK1R and nitric oxide synthase (NOS1) is still unknown for pancreatic acinar cells, but it is shown that neuronal NOS1 (nNOS) is activated by the Gβγ-subunit and activated tyrosine phosphatase SHP-2 in CHO cells. SHP-2 associated with the Gβ1 subunit, became activated, and then dephosphorylated nNOS through direct association [42].

Activated NOS1 cleaves L-arginine, forming L-citrulline and NO. NO then activates soluble guanylate cyclase [42], which produces cGMP from GMP. cGMP then (directly or via other components) activates a cytosolic ADP-ribosyl cyclase (CD38). This CD38 produces cyclic cADPr from NAD^+ [145]. cADPr then activates ryanodine receptor (RyR) in the endoplasmic reticulum, which then facilitates the transport of Ca^{2+} from the ER to the cytosol [146]. RyR is shown to be active in signaling in pancreatic acinar cells [43]. The Ca^{2+}-induced Ca^{2+} release (CICR) mechanism enhances calcium transport from ER to cytosol, and is also mediated by the ryanodine receptor [146].

Ryanodine receptors consist of three isoforms, RYR1, RYR2 and RYR3, and all three isoforms are expressed in rat pancreatic tissue, with RYR1 and RYR2 specifically found in pancreatic acinar cells [147]. Ca^{2+} is thus increased by both the IP3 and cADPr pathway, and it has been shown in mouse pancreatic acinar cells that CCK-induced Ca^{2+}-spiking can be mediated by both pathways, and that the pathway mediating the response is dependent on intracellular glucose levels: High glucose levels potentiates IP3-evoked Ca^{2+}-spiking, and low glucose levels potentiates cADPR-evoked Ca^{2+}-spiking [148].

Table 13

Module_NOS1					
Incoming			Defining		
components	UniProt accession	UniProt KB entry	components	UniProt accession	UniProt KB entry
SHP2	Q06124	PTN11_HUMAN	CD38	P28907	CD38_HUMAN
			NOS1	P29475	NOS1_HUMAN
			Guanylate cyclase	Q02846	GUC2D_HUMAN
			RYR1	P21817	RYR1_HUMAN
			RYR2	Q92736	RYR2_HUMAN
			RYR3	Q15413	RYR3_HUMAN
			cGK 1	Q13976	KGP1_HUMAN
			SHP2	Q06124	PTN11_HUMAN

PKA module

In the modular view, CCK1R module manifests positive influence on the PKA module (list of components in Table 14). CCK1R is coupled to G_S and CCK1R stimulation then activates adenylate cyclase [149]. Active adenylate cyclase converts ATP into cAMP which then activates cAMP-dependent protein kinase (cAPK)/PKA by releasing the catalytic subunits from the regulatory subunits [150]. PKA is a heterotetramer in its inactive form, with two regulatory subunits binding the catalytic subunits. Different subunits have different affinities for cAMP, generating holoenzymes (PKA type I or type II). Each regulatory subunit binds two cAMP molecules, releasing the catalytic subunits[151]. PKA then phosphorylates serine and threonine residues on specific substrate proteins both in the cytoplasm and in the nucleus [152]. The catalytic subunits of PKA translocates to the cell nucleus, where the transcript factor CREB is activated through phosphorylation [150].

Table 14

Module_PKA								
Incoming			Defining			Outgoing		
components	UniProt accession	UniProt KB entry	components	UniProt accession	UniProt KB entry	components	UniProt accession	UniProt KB entry
Gαs	P63092	GNAS2_HUMAN	Adenylate cyclase	Q08828	ADCY1_HUMAN	CREB1	P16220	CREB1_HUMAN
			AKAP	Q92667	AKAP1_HUMAN	IP3R	Q14643	ITPR1_HUMAN
			PKA-R2α	P13861	KAP2_HUMAN			
			PKA-Cα	P17612	KAPCA_HUMAN			
			PKA-Cβ	P22694	KAPCB_HUMAN			
			PDE	P54750	PDE1A_HUMAN			
			PKA-Cγ	P22612	KAPCG_HUMAN			
			PKA-R1α	P10644	KAP0_HUMAN			
			PKA-R1β	P31321	KAP1_HUMAN			
			PKA-R2β	P31323	KAPCG_HUMAN			

PKC module

In the comprehensive map, PKC module depicts activation of different members of the PKC family (detail list in Table 15). CCKR elicits DAG and IP3 production via PLC dependent mechanism by catalyzing PIP2. In CCK2R system, both PLCβ and PLCγ1 dependent IP3 production has been documented while in CCK1R system only PLCβ mediated DAG and IP3 production is reported. Activated IP3 then binds to IP3 receptor in the ER and triggers oscillation of Ca^{2+} from ER to cytosol. PKC superfamily has 3 different subfamilies: i) conventional PKCs, members of this family require both DAG and Ca^{2+} for activation. PKCα and PKCβ are the members of this family which are present in our model, ii) novel PKCs, members of this family are DAG responsive and Ca^{2+} unresponsive. PKCδ, PKCϵ, PKCη, and PKCθ are the members of this family, and iii) atypical PKCs, members of this family require neither DAG nor Ca^{2+} for activation. PKCζ is a member of this family which has been reported to play a role in CCKR signaling. Active PKCs are involved in activation of another serine/threonine kinase, protein kinase D (PRKD). Both CCK1R and CCK2R dependent activation of PRKD1 has been established while activation of PRKD2 is documented only downstream of CCK2R. The specific PKC isoforms associated with PRKD1 activation after CCK stimulation are PKCδ, PKCϵ, and PKCθ *[36, 140]*.

In human gastric cancer cells stably transfected with the CCKB/gastrin receptor , gastrin stimulates PRKD2 activation by PKC-α, -ε, and -η dependent phosphorylation of its residues (Sturany, Van Lint et al. 2001; Sturany, Van Lint et al. 2002; von Blume, Knippschild et al. 2007). Phosphorylation of PRKD2 at Ser244 within the zinc-finger domain by Casein Kinase (CK1)-δ and -ε promotes nuclear accumulation of PRKD2 in response to gastrin (von Blume, Knippschild et al. 2007).

Table 15

Module_PKC								
Incoming			Defining			Outgoing		
components	UniProt accession	UniProt KB entry	components	UniProt accession	UniProt KB entry	components	UniProt accession	UniProt KB entry
PKA-Cα	P17612	KAPCA_HUMAN	PRKD1	Q15139	KPCD1_HUMAN	SHC1	P29353	SHC1_HUMAN
PKA-Cβ	P22694	KAPCB_HUMAN	PKCδ	Q05655	KPCD_HUMAN	SRC	P12931	SRC_HUMAN
PKA-Cγ	P22612	KAPCG_HUMAN	PKCη	P24723	KPCL_HUMAN	RAF1	P04049	RAF1_HUMAN
Gαq	P29992	GNA11_HUMAN	PKCθ	Q04759	KPCT_HUMAN	MMP3	P08254	MMP3_HUMAN
RYR1	P21817	RYR1_HUMAN	PKCβ	P05771	KPCB_HUMAN	NFkB1	P19838	NFKB1_HUMAN
RYR3	Q15413	RYR2_HUMAN	PRKD2	Q9BZL6	KPCD2_HUMAN	RELA	Q04206	TF65_HUMAN
RYR2	Q92736	RYR3_HUMAN	PKCζ	Q05513	KPCZ_HUMAN	CCK1R	P32238	CCKAR_HUMAN
			PLCβ	Q9NQ66	PLCB1_HUMAN	RhoA	P61586	RHOA_HUMAN
			PKCε	Q02156	KPCE_HUMAN	YES1	P07947	YES_HUMAN
			IP3R	Q14643	ITPR1_HUMAN	LYN	P07948	LYN_HUMAN
			PLCγ1	P19174	PLCG1_HUMAN	HDAC7	Q8WUI4	HDAC7_HUMAN
			PKCα	P17252	KPCA_HUMAN	FAK2	Q14289	FAK2_HUMAN
						ARAF	P10398	ARAF_HUMAN
						HRAS	P01112	RASH_HUMAN

RAF1 module

RAF family constitutes three serine/threonine protein kinases, A-RAF, B-RAF, and C-RAF (RAF1) (Table 16). These protein kinases act as a regulatory link between membrane bound RAS-GTPase and MAPK cascade. CCK1R has been implicated to activate all three RAFs [36, 153] whereas only RAF1 is documented to be activated in response to gastrin [71]. RAF1 module mainly represents life cycle of the RAF1 activation. Active HRAS dissociates RAF1 from the RAF1-14-3-3 complex and then recruits it to the plasma membrane from the cytosol [154]. Sequential phosphorylation of serine/threonine/tyrosine (except S259) residues of the membrane attached RAF1 by different kinases results into an active RAF1. RAF1 activates MAPK1/3 cascade by triggering phosphorylation of MAP2K1/2 proteins. Active AKT1 inactivates RAF1 by phosphorylating its S259 residue [155].

Table 16

Module_RAF1								
Incoming			Defining			Outgoing		
components	UniProt accession	UniProt KB entry	components	UniProt accession	UniProt KB entry	components	UniProt accession	UniProt KB entry
HRAS	P01112	RASH_HUMAN	14-3-3β/α	P31946	1433B_HUMAN	MAP2K1	Q02750	MP2K1_HUMAN
PKCθ	Q04759	KPCT_HUMAN	ARAF	P10398	ARAF_HUMAN	MAP2K2	P36507	MP2K2_HUMAN
AKT1	P31749	AKT1_HUMAN	BRAF	P15056	BRAF_HUMAN			
			RAF1	P04049	RAF1_HUMAN			

Rho GTPase module

The Rho GTPase module represents components and reactions involved in the activation of members of the Rho GTPase family (Table 17). Members of this family include: RHOA, RAC1, and CDC42. Both gastrin and cholecystokinin can activate RHOA and RAC1, while only gastrin is reported to be involved in the activation of CDC42 [53]. CCK2R mediated activation of Rho GTPases (RHOA, RAC1 and CDC42) from the

inactive GDP-bound form to the active GTP-bound form is via $G\alpha_q$. Guanine exchange factors (GEFs), for example Leukemia-associated Rho guanine-nucleotide exchange factor (LARG) can serve as an effector for $G\alpha_q$ – coupled receptors [67] and GTPase-activating proteins (GAPs) hydrolyzes GTP to convert active GTP-bound form of Rho GTPases into inactive GDP-bound form Gastrin-stimulated RHOA acts through interaction with a serine/threonine kinase, ROCK whereas RAC1 and CDC42 acts through specific effector protein, PAK1 [53].

Table 17

Module_RhoGTPase								
Incoming			Defining			Outgoing		
components	UniProt accession	UniProt KB entry	components	UniProt accession	UniProt KB entry	components	UniProt accession	UniProt KB entry
Gαq	P29992	GNA11_HUMAN	PAK1	Q13153	PAK1_HUMAN	AKT1	P31749	AKT1_HUMAN
Gα13	Q14344	GNA13_HUMAN	ROCK1	Q13464	ROCK1_HUMAN	NFkB1	P19838	NFKB1_HUMAN
HRAS	P01112	RASH_HUMAN	CDC42	P60953	CDC42_HUMAN	RELA	Q04206	TF65_HUMAN
p85	P27986	P85A_HUMAN	RAC1	P63000	RAC1_HUMAN	BCL2L1	Q07817	B2CL1_HUMAN
p110	P42338	PK3CB_HUMAN	RHOA	P61586	RHOA_HUMAN	BAD	Q92934	BAD_HUMAN
			LARG	Q9NZN5	ARHGC_HUMAN	BAX	Q07812	BAX_HUMAN
			ARHGAP4	P98171	RHG04_HUMAN	BCL2	P10415	BCL2_HUMAN
			RGS2	P41220	RGS2_HUMAN	GSK3beta	P49841	GSK3B_HUMAN
						beta-catenin	P35222	CTNB1_HUMAN
						SNAI1	O95863	SNAI1_HUMAN

SRC module

This module represents the role of GRB2, SHC, and SRC proteins in the CCKR signaling (detail list in Table 18). Both SRC and SHC proteins are activated by CCKR. PKC dependent phosphorylation of SHC is reported for both CCK1 and CCK2 receptors [124, 156] while activation of SRC via PKC is documented only in response to gastrin [124]. The SHC-gene (SHC1) encodes three major isoforms of SHC, p46SHC, p52SHC, and p66SHC. Gastrin mediates time and dose dependent increase in tyrosine phosphorylation of p46 SHC and p52SHC isoforms of adaptor protein SHC1 in AR42J cells. Gastrin induced phosphorylation of SHC is dependent on SRC kinase [157] and PKC isoforms (PKC-α,-δ,-ε) [124]. Active SRC, SHC1 associate with ligand bound EGFR. Further, active EGFR associates with GRB2 at the membrane which then recruits SOS1 onto the membrane from the cytosol. SHC1 forms an active complex with GRB2-SOS1 which leads to the activation of HRAS-RAF1-MAPK cascade [123]. GRB2 and SRC are also involved in activation of the FAK2 cascade in response to cholecystokinin [158, 159].

Table 18

Module_SRC								
Incoming			Defining			Outgoing		
components	UniProt accession	UniProt KB entry	components	UniProt accession	UniProt KB entry	components	UniProt accession	UniProt KB entry
PKCα	P17252	KPCA_HUMAN	GRB2	P62993	GRB2_HUMAN	PLCγ1	P19174	PLCG1_HUMAN
PKCθ	Q04759	KPCT_HUMAN	SRC	P12931	SRC_HUMAN	FAK1	Q05397	FAK1_HUMAN
Gαq	P29992	GNA11_HUMAN	SOS1	Q07889	SOS1_HUMAN	FAK2	Q14289	FAK2_HUMAN
PKCδ	Q05655	KPCD_HUMAN	HRAS	P01112	RASH_HUMAN	RAF1	P04049	RAF1_HUMAN
PKCε	Q02156	KPCE_HUMAN	CSK	P41240	CSK_HUMAN	ARAF	P10398	ARAF_HUMAN
IRS1	P35568	IRS1_HUMAN	SHC2	P98077	SHC2_HUMAN	BRAF	P15056	BRAF_HUMAN
FAK2	Q14289	FAK2_HUMAN	SHC3	Q92529	SHC3_HUMAN	RhoA	P61586	RHOA_HUMAN
			SHC1	P29353	SHC1_HUMAN	CDC42	P60953	CDC42_HUMAN
						AKT1	P31749	AKT1_HUMAN
						MAP3K11	Q16584	M3K11_HUMAN

References

1. Rehfeld JF: **The New Biology of Gastrointestinal Hormones**. *Physiological reviews* 1998, **78**(4):1087-1108.
2. Watson SA, Grabowska AM, El-Zaatari M, Takhar A: **Gastrin - active participant or bystander in gastric carcinogenesis?** *Nature reviews Cancer* 2006, **6**(12):936-946.
3. Little TJ, Horowitz M, Feinle-Bisset C: **Role of cholecystokinin in appetite control and body weight regulation**. *Obesity Reviews* 2005, **6**(4):297-306.
4. Saluja aK, Saluja M, Printz H, Zavertnik a, Sengupta a, Steer ML: **Experimental pancreatitis is mediated by low-affinity cholecystokinin receptors that inhibit digestive enzyme secretion.** *Proceedings of the National Academy of Sciences of the United States of America* 1989, **86**:8968-8971.
5. Gukovsky I, Cheng JH, Nam KJ, Lee OT, Lugea A, Fischer L, Penninger JM, Pandol SJ, Gukovskaya AS: **Phosphatidylinositide 3-kinase gamma regulates key pathologic responses to cholecystokinin in pancreatic acinar cells.** *Gastroenterology* 2004, **126**:554-566.
6. Dabrowski A, Grady T, Logsdon CD, Williams Ja: **Jun kinases are rapidly activated by cholecystokinin in rat pancreas both in vitro and in vivo.** *The Journal of biological chemistry* 1996, **271**:5686-5690.
7. Gibbs J, Young RC, Smith GP: **Cholecystokinin decreases food intake in rats.** *Journal of comparative and physiological psychology* 1973, **84**:488-495.
8. Witkamp RF: **Current and future drug targets in weight management.** *Pharmaceutical research* 2011, **28**:1792-1818.
9. Varga G, Bálint A, Burghardt B, D'Amato M: **Involvement of endogenous CCK and CCK1 receptors in colonic motor function.** *British journal of pharmacology* 2004, **141**:1275-1284.
10. Dufresne M, Seva C, Fourmy D: **Cholecystokinin and gastrin receptors**. *Physiological reviews* 2006, **86**(3):805-847.
11. Cawston EE, Miller LJ: **Therapeutic potential for novel drugs targeting the type 1 cholecystokinin receptor**. *British journal of pharmacology* 2010, **159**(5):1009-1021.
12. Konturek PC, Konturek SJ, Brzozowski T: **Helicobacter pylori infection in gastric cancerogenesis**. *Journal of physiology and pharmacology : an official journal of the Polish Physiological Society* 2009, **60**(3):3-21.

13. Matysiak-Budnik T, Mégraud F: **Helicobacter pylori infection and gastric cancer**. *European Journal of Cancer* 2006, **42**(6):708-716.
14. Noble F, Wank SA, Crawley JN, Bradwejn J, Seroogy KB, Hamon M, Roques BP: **International Union of Pharmacology. XXI. Structure, Distribution, and Functions of Cholecystokinin Receptors**. *Pharmacological Reviews* 1999, **51**(4):745-781.
15. Christophe J: **Pancreatic tumoral cell line AR42J: an amphicrine model**. *American Journal of Physiology - Gastrointestinal and Liver Physiology* 1994, **266**(6):G963-G971.
16. Ochsner SA, Watkins CM, LaGrone BS, Steffen DL, McKenna NJ: **Research Resource: Tissue-Specific Transcriptomics and Cistromics of Nuclear Receptor Signaling: A Web Research Resource**. *Molecular Endocrinology* 2010, **24**(10):2065-2069.
17. Diehl CJ, Barish GD, Downes M, Chou MY, Heinz S, Glass CK, Evans RM, Witztum JL: **Research Resource: Comparative Nuclear Receptor Atlas: Basal and Activated Peritoneal B-1 and B-2 Cells**. *Molecular Endocrinology* 2011, **25**(3):529-545.
18. Lee MJ, Ye AS, Gardino AK, Heijink AM, Sorger PK, MacBeath G, Yaffe MB: **Sequential Application of Anticancer Drugs Enhances Cell Death by Rewiring Apoptotic Signaling Networks**. *Cell* 2012, **149**(4):780-794.
19. Kaizu K, Ghosh S, Matsuoka Y, Moriya H, Shimizu-Yoshida Y, Kitano H: **A comprehensive molecular interaction map of the budding yeast cell cycle**. *Mol Syst Biol* 2010, **6**.
20. Gloaguen P, Crépieux P, Heitzler D, Poupon A, Reiter E: **Mapping the follicle-stimulating hormone-induced signalling networks**. *Frontiers in Endocrinology* 2011, **2**.
21. Oda K, Matsuoka Y, Funahashi A, Kitano H: **A comprehensive pathway map of epidermal growth factor receptor signaling**. *Mol Syst Biol* 2005, **1**:2005 0010.
22. Calzone L, Gelay A, Zinovyev A, Radvanyi F, Barillot E: **A comprehensive modular map of molecular interactions in RB/E2F pathway**. *Mol Syst Biol* 2008, **4**.
23. Caron E, Ghosh S, Matsuoka Y, Ashton-Beaucage D, Therrien M, Lemieux S, Perreault C, Roux PP, Kitano H: **A comprehensive map of the mTOR signaling network**. *Mol Syst Biol* 2010, **6**.
24. Fink MY, Pincas H, Choi SG, Nudelman G, Sealfon SC: **Research Resource: Gonadotropin-Releasing Hormone Receptor-Mediated Signaling Network in L beta T2 Cells: A Pathway-Based Web-Accessible Knowledgebase**. *Molecular Endocrinology* 2010, **24**(9):1863-1871.
25. Mizuno S, Iijima R, Ogishima S, Kikuchi M, Matsuoka Y, Ghosh S, Miyamoto T, Miyashita A, Kuwano R, Tanaka H: **AlzPathway: a comprehensive map of signaling pathways of Alzheimer's disease**. *BMC Syst Biol* 2012, **6**:52.
26. Patil S, Pincas H, Seto J, Nudelman G, Nudelman I, Sealfon SC: **Signaling network of dendritic cells in response to pathogens: a community-input supported knowledgebase**. *BMC systems biology* 2010, **4**:137.
27. Raza S, McDerment N, Lacaze PA, Robertson K, Watterson S, Chen Y, Chisholm M, Eleftheriadis G, Monk S, O'Sullivan M *et al*: **Construction of a large scale integrated map of macrophage pathogen recognition and effector systems**. *BMC systems biology* 2010, **4**:63.
28. Zinovyev A, Viara E, Calzone L, Barillot E: **BiNoM: a Cytoscape plugin for manipulating and analyzing biological networks**. *Bioinformatics* 2008, **24**(6):876-877.
29. Glaab E, Baudot A, Krasnogor N, Valencia A: **Extending pathways and processes using molecular interaction networks to analyse cancer genome data**. *BMC bioinformatics* 2010, **11**:597.
30. Funahashi A, Morohashi M, Kitano H, Tanimura N: **CellDesigner: a process diagram editor for gene-regulatory and biochemical networks**. *BIOSILICO* 2003, **1**(5):159-162.
31. Matsuoka Y, Ghosh S, Kikuchi N, Kitano H: **Payao: a community platform for SBML pathway model curation**. *Bioinformatics* 2010, **26**(10):1381-1383.

32. Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, Caudy M, Garapati P, Gopinath G, Jassal B *et al*: **Reactome: a database of reactions, pathways and biological processes**. *Nucleic Acids Res* 2011, **39**(Database issue):D691-697.
33. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M: **KEGG for integration and interpretation of large-scale molecular data sets**. *Nucleic Acids Research* 2012, **40**(D1):D109-D114.
34. Dufresne M, Seva C, Fourmy D: **Cholecystokinin and gastrin receptors.** *Physiological reviews* 2006, **86**:805-847.
35. Paillasse MR, de Medina P, Amouroux G, Mhamdi L, Poirot M, Silvente-Poirot S: **Signaling through cholesterol esterification: a new pathway for the cholecystokinin 2 receptor involved in cell growth and invasion**. *Journal of Lipid Research* 2009, **50**(11):2203-2211.
36. Sancho V, Berna MJ, Thill M, Jensen RT: **PKCθ activation in pancreatic acinar cells by gastrointestinal hormones/neurotransmitters and growth factors is needed for stimulation of numerous important cellular signaling cascades**. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* 2011, **1813**(12):2145-2156.
37. Quattrone A, Dewaele B, Wozniak A, Bauters M, Vanspauwen V, Floris G, Schoffski P, Chibon F, Coindre JM, Sciot R *et al*: **Promoting role of cholecystokinin 2 receptor (CCK2R) in gastrointestinal stromal tumours pathogenesis**. *The Journal of pathology* 2012.
38. Wu V, Yang M, McRoberts Ja, Ren J, Seensalu R, Zeng N, Dagrag M, Birnbaumer M, Walsh JH: **First intracellular loop of the human cholecystokinin-A receptor is essential for cyclic AMP signaling in transfected HEK-293 cells.** *The Journal of biological chemistry* 1997, **272**:9037-9042.
39. Sabbatini ME, Bi Y, Ji B, Ernst Sa, Williams Ja: **CCK activates RhoA and Rac1 differentially through Galpha13 and Galphaq in mouse pancreatic acini.** *American journal of physiology Cell physiology* 2010, **298**:C592-601.
40. Le Page SL, Bi Y, Williams Ja: **CCK-A receptor activates RhoA through G alpha 12/13 in NIH3T3 cells.** *American journal of physiology Cell physiology* 2003, **285**:C1197-1206.
41. Moustafa A, Sakamoto KQ, Habara Y: **A fundamental role for NO-PLC signaling pathway in mediating intracellular Ca2+ oscillation in pancreatic acini.** *Nitric oxide : biology and chemistry / official journal of the Nitric Oxide Society* 2011, **24**:139-150.
42. Cordelier P, Estève JP, Rivard N, Marletta M, Vaysse N, Susini C, Buscail L: **The activation of neuronal NO synthase is mediated by G-protein betagamma subunit and the tyrosine phosphatase SHP-2.** *The FASEB journal : official publication of the Federation of American Societies for Experimental Biology* 1999, **13**:2037-2050.
43. Thorn P, Gerasimenko O, Petersen OH: **Cyclic ADP-ribose regulation of ryanodine receptors involved in agonist evoked cytosolic Ca2+ oscillations in pancreatic acinar cells.** *The EMBO journal* 1994, **13**:2038-2043.
44. Cosker F, Cheviron N, Yamasaki M, Menteyne A, Lund FE, Moutin M-J, Galione A, Cancela J-M: **The ecto-enzyme CD38 is a nicotinic acid adenine dinucleotide phosphate (NAADP) synthase that couples receptor activation to Ca2+ mobilization from lysosomes in pancreatic acinar cells.** *The Journal of biological chemistry* 2010, **285**:38251-38259.
45. Calcraft PJ, Ruas M, Pan Z, Cheng X, Arredouani A, Hao X, Tang J, Rietdorf K, Teboul L, Chuang K-t *et al*: **NAADP mobilizes calcium from acidic organelles through two-pore channels.** *Nature* 2009, **459**:596-600.
46. Sinclair NF, Ai W, Raychowdhury R, Bi M, Wang TC, Koh TJ, McLaughlin JT: **Gastrin regulates the heparin-binding epidermal-like growth factor promoter via a PKC/EGFR-dependent mechanism**. *American Journal of Physiology - Gastrointestinal and Liver Physiology* 2004, **286**(6):G992-G999.
47. Miyazaki Y, Shinomura Y, Tsutsui S, Zushi S, Higashimoto Y, Kanayama S, Higashiyama S, Taniguchi N, Matsuzawa Y: **Gastrin induces heparin-binding epidermal growth factor-**

like growth factor in rat gastric epithelial cells transfected with gastrin receptor. *Gastroenterology* 1999, **116**(1):78-89.
48. Sturany S, Van Lint J, Gilchrist A, Vandenheede JR, Adler G, Seufferlein T: **Mechanism of Activation of Protein Kinase D2(PKD2) by the CCKB/Gastrin Receptor**. *Journal of Biological Chemistry* 2002, **277**(33):29431-29436.
49. Yassin RR, Little KM: **Early signalling mechanism in colonic epithelial cell response to gastrin**. *The Biochemical journal* 1995, **311 (Pt 3)**:945-950.
50. He H, Shulkes A, Baldwin GS: **PAK1 interacts with beta-catenin and is required for the regulation of the beta-catenin signalling pathway by gastrins**. *Biochimica et biophysica acta* 2008, **1783**(10):1943-1954.
51. Mishra P, Senthivinayagam S, Rana A, Rana B: **Glycogen Synthase Kinase-3beta regulates Snail and beta-catenin during gastrin-induced migration of gastric cancer cells**. *J Mol Signal* 2010, **5**:9.
52. He H, Baldwin GS: **Rho GTPases and p21-activated kinase in the regulation of proliferation and apoptosis by gastrins**. *Int J Biochem Cell Biol* 2008, **40**(10):2018-2022.
53. He H, Yim M, Liu KH, Cody SC, Shulkes A, Baldwin GS: **Involvement of G proteins of the Rho family in the regulation of Bcl-2-like protein expression and caspase 3 activation by Gastrins**. *Cellular signalling* 2008, **20**(1):83-93.
54. Guo Y-S, Cheng J-Z, Jin G-F, Gutkind JS, Hellmich MR, Townsend CM: **Gastrin stimulates cyclooxygenase-2 expression in intestinal epithelial cells through multiple signaling pathways. Evidence for involvement of ERK5 kinase and transactivation of the epidermal growth factor receptor.** *The Journal of biological chemistry* 2002, **277**:48755-48763.
55. von Blume J, Knippschild U, Dequiedt F, Giamas G, Beck A, Auer A, Van Lint J, Adler G, Seufferlein T: **Phosphorylation at Ser244 by CK1 determines nuclear localization and substrate targeting of PKD2**. *The EMBO journal* 2007, **26**(22):4619-4633.
56. Seva C, Kowalski-Chauvel A, Daulhac L, Barthez C, Vaysse N, Pradayrol L: **Wortmannin-Sensitive Activation of p70S6-Kinase and MAP-Kinase by the G Protein-Coupled Receptor, G/CCKB**. *Biochemical and biophysical research communications* 1997, **238**(1):202-206.
57. Kikani CK, Dong LQ, Liu F: **"New"-clear functions of PDK1: beyond a master kinase in the cytosol?** *Journal of cellular biochemistry* 2005, **96**(6):1157-1162.
58. Pradeep A, Sharma C, Sathyanarayana P, Albanese C, Fleming JV, Wang TC, Wolfe MM, Baker KM, Pestell RG, Rana B: **Gastrin-mediated activation of cyclin D1 transcription involves beta-catenin and CREB pathways in gastric cancer cells**. *Oncogene* 2004, **23**(20):3689-3699.
59. Steigedal TS, Bruland T, Misund K, Thommesen L, Laegreid A: **Inducible cAMP early repressor suppresses gastrin-mediated activation of cyclin D1 and c-fos gene expression**. *Am J Physiol Gastrointest Liver Physiol* 2007, **292**(4):G1062-1069.
60. Bierkamp C, Kowalski-Chauvel A, Dehez S, Fourmy D, Pradayrol L, Seva C: **Gastrin mediated cholecystokinin-2 receptor activation induces loss of cell adhesion and scattering in epithelial MDCK cells**. *Oncogene* 2002, **21**(50):7656-7670.
61. Mishra P, Senthivinayagam S, Rangasamy V, Sondarva G, Rana B: **Mixed Lineage Kinase-3/JNK1 Axis Promotes Migration of Human Gastric Cancer Cells following Gastrin Stimulation**. *Molecular Endocrinology* 2010, **24**(3):598-607.
62. Fjeldbo CS, Bakke I, Erlandsen SE, Holmseth J, Laegreid A, Sandvik AK, Thommesen L, Bruland T: **Gastrin upregulates the prosurvival factor secretory clusterin in adenocarcinoma cells and in oxyntic mucosa of hypergastrinemic rats**. *Am J Physiol Gastrointest Liver Physiol* 2012, **302**(1):G21-33.

63. Ramamoorthy S, Stepan V, Todisco A: **Intracellular mechanisms mediating the anti-apoptotic action of gastrin**. *Biochemical and biophysical research communications* 2004, **323**(1):44-48.
64. Durmu, #351, Tekir S, #220, mit P, Eren Toku A, lgen K, #214: **Reconstruction of Protein-Protein Interaction Network of Insulin Signaling in Homo Sapiens**. *Journal of Biomedicine and Biotechnology* 2010, **2010**.
65. Paris L, Bazzoni G: **The protein interaction network of the epithelial junctional complex: a system-level analysis**. *Mol Biol Cell* 2008, **19**(12):5409-5421.
66. Maere S, Heymans K, Kuiper M: **BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks**. *Bioinformatics* 2005, **21**(16):3448-3449.
67. Booden MA, Siderovski DP, Der CJ: **Leukemia-associated Rho guanine nucleotide exchange factor promotes G alpha q-coupled activation of RhoA**. *Mol Cell Biol* 2002, **22**(12):4053-4061.
68. Cinar B, Fang PK, Lutchman M, Di Vizio D, Adam RM, Pavlova N, Rubin MA, Yelick PC, Freeman MR: **The pro-apoptotic kinase Mst1 and its caspase cleavage products are direct inhibitors of Akt1**. *Embo Journal* 2007, **26**(21):4523-4534.
69. Liao JC, Boscolo R, Yang YL, Tran LM, Sabatti C, Roychowdhury VP: **Network component analysis: reconstruction of regulatory signals in biological systems**. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**(26):15522-15527.
70. Hocker M: **Molecular mechanisms of gastrin-dependent gene regulation**. *Annals of the New York Academy of Sciences* 2004, **1014**:97-109.
71. Seufferlein T, Withers D, Broad S, Herget T, Walsh J, Rozengurt E: **The human CCKB/gastrin receptor transfected into rat1 fibroblasts mediates activation of MAP kinase, p74raf-1 kinase, and mitogenesis**. *Cell Growth Differ* 1995, **6**(4):383-393.
72. Hur EM, Zhou FQ: **GSK3 signalling in neural development**. *Nature reviews Neuroscience* 2010, **11**(8):539-551.
73. Grimes CA, Jope RS: **CREB DNA binding activity is inhibited by glycogen synthase kinase-3 beta and facilitated by lithium**. *J Neurochem* 2001, **78**(6):1219-1232.
74. Wu GY, Deisseroth K, Tsien RW: **Activity-dependent CREB phosphorylation: convergence of a fast, sensitive calmodulin kinase pathway and a slow, less sensitive mitogen-activated protein kinase pathway**. *Proc Natl Acad Sci U S A* 2001, **98**(5):2808-2813.
75. Delghandi MP, Johannessen M, Moens U: **The cAMP signalling pathway activates CREB through PKA, p38 and MSK1 in NIH 3T3 cells**. *Cellular signalling* 2005, **17**(11):1343-1351.
76. Stepan V, Ramamoorthy S, Pausawasdi N, Logsdon CD, Askari FK, Todisco A: **Role of small GTP binding proteins in the growth-promoting and antiapoptotic actions of gastrin**. *Am J Physiol Gastrointest Liver Physiol* 2004, **287**(3):G715-725.
77. Sheppard KA, Rose DW, Haque ZK, Kurokawa R, McInerney E, Westin S, Thanos D, Rosenfeld MG, Glass CK, Collins T: **Transcriptional activation by NF-kappaB requires multiple coactivators**. *Mol Cell Biol* 1999, **19**(9):6367-6378.
78. Na SY, Lee SK, Han SJ, Choi HS, Im SY, Lee JW: **Steroid receptor coactivator-1 interacts with the p50 subunit and coactivates nuclear factor kappaB-mediated transactivations**. *The Journal of biological chemistry* 1998, **273**(18):10831-10834.
79. El-Asmar B, Giner XC, Tremblay JJ: **Transcriptional cooperation between NF-kappaB p50 and CCAAT/enhancer binding protein beta regulates Nur77 transcription in Leydig cells**. *Journal of molecular endocrinology* 2009, **42**(2):131-138.
80. Dooher JE, Paz-Priel I, Houng S, Baldwin AS, Jr., Friedman AD: **C/EBPalpha, C/EBPalpha oncoproteins, or C/EBPbeta preferentially bind NF-kappaB p50 compared with p65, focusing therapeutic targeting on the C/EBP:p50 interaction**. *Molecular cancer research : MCR* 2011, **9**(10):1395-1405.

81. Huang YS, Shih HM: **Daxx positively modulates beta-catenin/TCF4-mediated transcriptional potential**. *Biochemical and biophysical research communications* 2009, **386**(4):762-768.
82. Lukas J, Mazna P, Valenta T, Doubravska L, Pospichalova V, Vojtechova M, Fafilek B, Ivanek R, Plachy J, Novak J *et al*: **Dazap2 modulates transcription driven by the Wnt effector TCF-4**. *Nucleic Acids Research* 2009, **37**(9):3007-3020.
83. Valenta T, Lukas J, Doubravska L, Fafilek B, Korinek V: **HIC1 attenuates Wnt signaling by recruitment of TCF-4 and beta-catenin to the nuclear bodies**. *Embo Journal* 2006, **25**(11):2326-2337.
84. Jun Cao J-PY, Chao-Hong Liu, Lan Zhou, Hong-Gang Yu: **Effects of gastrin 17 on β-catenin/Tcf-4 pathway in Colo320WT colon cancer cells**. *World journal of gastroenterology : WJG* 2006 **12**(46):7482-7487.
85. Clarke PA, Dickson JH, Harris JC, Grabowska A, Watson SA: **Gastrin enhances the angiogenic potential of endothelial cells via modulation of heparin-binding epidermal-like growth factor**. *Cancer Res* 2006, **66**(7):3504-3512.
86. Trulsson LM, Gasslander T, Svanvik J: **Cholecystokinin-8-induced hypoplasia of the rat pancreas: influence of nitric oxide on cell proliferation and programmed cell death**. *Basic & clinical pharmacology & toxicology* 2004, **95**(4):183-190.
87. Leach SM, Tipney H, Feng W, Baumgartner WA, Kasliwal P, Schuyler RP, Williams T, Spritz RA, Hunter L: **Biomedical discovery acceleration, with applications to craniofacial development**. *PLoS Comput Biol* 2009, **5**(3):e1000215.
88. Gitter A, Carmi M, Barkai N, Bar-Joseph Z: **Linking the signaling cascades and dynamic regulatory networks controlling stress responses**. *Genome research* 2012.
89. Tiger CF, Krause F, Cedersund G, Palmer R, Klipp E, Hohmann S, Kitano H, Krantz M: **A framework for mapping, visualisation and automatic model creation of signal-transduction networks**. *Mol Syst Biol* 2012, **8**:578.
90. Novere NL, Hucka M, Mi H, Moodie S, Schreiber F, Sorokin A, Demir E, Wegner K, Aladjem MI, Wimalaratne SM *et al*: **The Systems Biology Graphical Notation**. *Nat Biotech* 2009, **27**(8):735-741.
91. Kitano H, Funahashi A, Matsuoka Y, Oda K: **Using process diagrams for the graphical representation of biological networks**. *Nature biotechnology* 2005, **23**(8):961-966.
92. Novere NL, Finney A, Hucka M, Bhalla US, Campagne F, Collado-Vides J, Crampin EJ, Halstead M, Klipp E, Mendes P *et al*: **Minimum information requested in the annotation of biochemical models (MIRIAM)**. *Nat Biotech* 2005, **23**(12):1509-1515.
93. Bader GD, Cary MP, Sander C: **Pathguide: a Pathway Resource List**. *Nucleic Acids Research*, **34**(suppl 1):D504-D506.
94. Frisch M, Klocke B, Haltmeier M, Frech K: **LitInspector: literature and signal transduction pathway mining in PubMed abstracts**. *Nucleic Acids Research* 2009, **37**(suppl 2):W135-W140.
95. Hoffmann R, Valencia A: **A gene network for navigating the literature**. *Nature genetics* 2004, **36**(7):664.
96. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T: **Cytoscape 2.8: new features for data integration and network visualization**. *Bioinformatics* 2011, **27**(3):431-432.
97. Assenov Y, Ramirez F, Schelhorn SE, Lengauer T, Albrecht M: **Computing topological parameters of biological networks**. *Bioinformatics* 2008, **24**(2):282-284.
98. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: A software environment for integrated models of biomolecular interaction networks**. *Genome research* 2003, **13**(11):2498-2504.
99. Souiai O, Becker E, Prieto C, Benkahla A, De Las Rivas J, Brun C: **Functional Integrative Levels in the Human Interactome Recapitulate Organ Organization**. *PLoS ONE* 2011, **6**(7).

100. Essaghir A, Toffalini F, Knoops L, Kallin A, van Helden J, Demoulin JB: **Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data**. *Nucleic Acids Research* 2010, **38**(11).
101. Kim M, Nozu F, Kusama K, Imawari M: **Cholecystokinin stimulates the recruitment of the Src-RhoA-phosphoinositide 3-kinase pathway by Vav-2 downstream of G(alpha 13) in pancreatic acini**. *Biochemical and biophysical research communications* 2006, **339**(1):271-276.
102. Daulhac L, Kowalski-Chauvel A, Pradayrol L, Vaysse N, Seva C: **Gastrin stimulates the formation of a p60Src/p125FAK complex upstream of the phosphatidylinositol 3-kinase signaling pathway**. *FEBS letters* 1999, **445**(2–3):251-255.
103. Kowalski-Chauvel A, Pradayrol L, Vaysse N, Seva C: **Gastrin stimulates tyrosine phosphorylation of insulin receptor substrate 1 and its association with Grb2 and the phosphatidylinositol 3-kinase**. *The Journal of biological chemistry* 1996, **271**(42):26356-26361.
104. Daulhac L, Kowalski-Chauvel A, Pradayrol L, Vaysse N, Seva C: **Src-family Tyrosine Kinases in Activation of ERK-1 and p85/p110-phosphatidylinositol 3-Kinase by G/CCKBReceptors**. *Journal of Biological Chemistry* 1999, **274**(29):20657-20663.
105. Ferrand A, Kowalski-Chauvel A, Bertrand C, Pradayrol L, Fourmy D, Dufresne M, Seva C: **Involvement of JAK2 upstream of the PI 3-kinase in cell–cell adhesion regulation by gastrin**. *Experimental Cell Research* 2004, **301**(2):128-138.
106. Todisco A, Ramamoorthy S, Witham T, Pausawasdi N, Srinivasan S, Dickinson CJ, Askari FK, Krametter D: **Molecular mechanisms for the antiapoptotic action of gastrin**. *Am J Physiol Gastrointest Liver Physiol* 2001, **280**(2):G298-307.
107. Brader S, Eccles SA: **Phosphoinositide 3-kinase signalling pathways in tumor progression, invasion and angiogenesis**. *Tumori* 2004, **90**(1):2-8.
108. Subramaniam D, Ramalingam S, May R, Dieckgraefe BK, Berg DE, Pothoulakis C, Houchen CW, Wang TC, Anant S: **Gastrin-Mediated Interleukin-8 and Cyclooxygenase-2 Gene Expression: Differential Transcriptional and Posttranscriptional Mechanisms**. *Gastroenterology* 2008, **134**(4):1070-1082.
109. Minden A, Lin A, Smeal T, Derijard B, Cobb M, Davis R, Karin M: **c-Jun N-terminal phosphorylation correlates with activation of the JNK subgroup but not the ERK subgroup of mitogen-activated protein kinases**. *Mol Cell Biol* 1994, **14**(10):6683-6688.
110. Coronella-Wood J, Terrand J, Sun H, Chen QM: **c-Fos phosphorylation induced by H2O2 prevents proteasomal degradation of c-Fos in cardiomyocytes**. *The Journal of biological chemistry* 2004, **279**(32):33567-33574.
111. Livingstone C, Patel G, Jones N: **Atf-2 Contains a Phosphorylation-Dependent Transcriptional Activation Domain**. *Embo Journal* 1995, **14**(8):1785-1797.
112. Gupta S, Campbell D, Derijard B, Davis RJ: **Transcription Factor Atf2 Regulation by the Jnk Signal-Transduction Pathway**. *Science* 1995, **267**(5196):389-393.
113. Pritchard DM, Berry D, Przemeck SMC, Campbell F, Edwards SW, Varro A: **Gastrin increases mcl-1 expression in type I gastric carcinoid tumors and a gastric epithelial cell line that expresses the CCK-2 receptor**. *American Journal of Physiology - Gastrointestinal and Liver Physiology* 2008, **295**(4):G798-G805.
114. Michels J, Johnson PWM, Packham G: **Mcl-1**. *The International Journal of Biochemistry & Cell Biology* 2005, **37**(2):267-271.
115. Schürmann A, Mooney AF, Sanders LC, Sells MA, Wang HG, Reed JC, Bokoch GM: **p21-Activated Kinase 1 Phosphorylates the Death Agonist Bad and Protects Cells from Apoptosis**. *Molecular and Cellular Biology* 2000, **20**(2):453-461.
116. Ozcelebi F, Rao RV, Holicky E, Madden BJ, McCormick DJ, Miller LJ: **Phosphorylation of cholecystokinin receptors expressed on Chinese hamster ovary cells. Similarities and**

differences relative to native pancreatic acinar cell receptors. *The Journal of biological chemistry* 1996, **271**(7):3750-3755.
117. Rao RV, Roettger BF, Hadac EM, Miller LJ: **Roles of cholecystokinin receptor phosphorylation in agonist-stimulated desensitization of pancreatic acinar cells and receptor-bearing Chinese hamster ovary cholecystokinin receptor cells**. *Mol Pharmacol* 1997, **51**(2):185-192.
118. Pohl M, Silvente-Poirot S, Pisegna JR, Tarasova NI, Wank SA: **Ligand-induced internalization of cholecystokinin receptors. Demonstration of the importance of the carboxyl terminus for ligand-induced internalization of the rat cholecystokinin type B receptor but not the type A receptor**. *The Journal of biological chemistry* 1997, **272**(29):18179-18184.
119. Tarasova NI, Stauber RH, Choi JK, Hudson EA, Czerwinski G, Miller JL, Pavlakis GN, Michejda CJ, Wank SA: **Visualization of G Protein-coupled Receptor Trafficking with the Aid of the Green Fluorescent Protein**. *Journal of Biological Chemistry* 1997, **272**(23):14817-14824.
120. Suzuki M, Raab G, Moses MA, Fernandez CA, Klagsbrun M: **Matrix Metalloproteinase-3 Releases Active Heparin-binding EGF-like Growth Factor by Cleavage at a Specific Juxtamembrane Site**. *Journal of Biological Chemistry* 1997, **272**(50):31730-31737.
121. Sakaguchi K, Okabayashi Y, Kido Y, Kimura S, Matsumura Y, Inushima K, Kasuga M: **Shc phosphotyrosine-binding domain dominantly interacts with epidermal growth factor receptors and mediates Ras activation in intact cells**. *Mol Endocrinol* 1998, **12**(4):536-543.
122. Okutani T, Okabayashi Y, Kido Y, Sugimoto Y, Sakaguchi K, Matuoka K, Takenawa T, Kasuga M: **Grb2/Ash binds directly to tyrosines 1068 and 1086 and indirectly to tyrosine 1148 of activated human epidermal growth factor receptors in intact cells**. *The Journal of biological chemistry* 1994, **269**(49):31310-31314.
123. Seva C, Kowalski-Chauvel A, Blanchet JS, Vaysse N, Pradayrol L: **Gastrin induces tyrosine phosphorylation of Shc proteins and their association with the Grb2/Sos complex**. *FEBS letters* 1996, **378**(1):74-78.
124. Daulhac L, Kowalski-Chauvel A, Pradayrol L, Vaysse N, Seva C: **Ca2+ and protein kinase C-dependent mechanisms involved in gastrin-induced Shc/Grb2 complex formation and P44-mitogen-activated protein kinase activation**. *The Biochemical journal* 1997, **325 (Pt 2)**:383-389.
125. Magnan R, Masri B, Escrieut C, Foucaud M, Cordelier P, Fourmy D: **Regulation of Membrane Cholecystokinin-2 Receptor by Agonists Enables Classification of Partial Agonists as Biased Agonists**. *Journal of Biological Chemistry* 2011, **286**(8):6707-6719.
126. Yu H-G, Schrader H, Otte J-M, Schmidt WE, Schmitz F: **Rapid tyrosine phosphorylation of focal adhesion kinase, paxillin, and p130Cas by gastrin in human colon cancer cells**. *Biochemical Pharmacology* 2004, **67**(1):135-146.
127. Rozengurt E, Walsh JH: **GASTRIN, CCK, SIGNALING, AND CANCER**. *Annual review of physiology* 2001, **63**(1):49-76.
128. Dehez S, Bierkamp C, Kowalski-Chauvel A, Daulhac L, Escrieut C, Susini C, Pradayrol L, Fourmy D, Seva C: **c-Jun NH2-terminal Kinase Pathway in Growth-promoting Effect of the G Protein-coupled Receptor Cholecystokinin B Receptor: A Protein Kinase C/Src-dependent-Mechanism**. *Cell Growth Differ* 2002, **13**(8):375-385.
129. Tapia Ja, Ferris Ha, Jensen RT, García LJ: **Cholecystokinin activates PYK2/CAKbeta by a phospholipase C-dependent mechanism and its association with the mitogen-activated protein kinase signaling pathway in pancreatic acinar cells.** *The Journal of biological chemistry* 1999, **274**:31261-31271.
130. Pace A, García-Marin LJ, Tapia Ja, Bragado MJ, Jensen RT: **Phosphospecific site tyrosine phosphorylation of p125FAK and proline-rich kinase 2 is differentially regulated by**

cholecystokinin receptor type A activation in pancreatic acini. *The Journal of biological chemistry* 2003, **278**:19008-19016.
131. Mitra SK, Hanson Da, Schlaepfer DD: **Focal adhesion kinase: in command and control of cell motility.** *Nature reviews Molecular cell biology* 2005, **6**:56-68.
132. Lipinski CA, Loftus JC: **Targeting Pyk2 for therapeutic intervention.** *Expert opinion on therapeutic targets* 2010, **14**:95-108.
133. Blaukat A, Ivankovic-Dikic I, Grönroos E, Dolfi F, Tokiwa G, Vuori K, Dikic I: **Adaptor proteins Grb2 and Crk couple Pyk2 with activation of specific mitogen-activated protein kinase cascades.** *The Journal of biological chemistry* 1999, **274**:14893-14901.
134. Andreolotti AG, Bragado MJ, Tapia Ja, Jensen RT, Garcia-Marin LJ: **Adapter protein CRKII signaling is involved in the rat pancreatic acini response to reactive oxygen species.** *Journal of cellular biochemistry* 2006, **97**:359-367.
135. Anjum R, Blenis J: **The RSK family of kinases: emerging roles in cellular signalling**. *Nat Rev Mol Cell Biol* 2008, **9**(10):747-758.
136. Hocker M, Raychowdhury R, Plath T, Wu H, O'Connor DT, Wiedenmann B, Rosewicz S, Wang TC: **Sp1 and CREB mediate gastrin-dependent regulation of chromogranin A promoter activity in gastric carcinoma cells**. *The Journal of biological chemistry* 1998, **273**(51):34000-34007.
137. Guo Y-S, Cheng J-Z, Jin G-F, Gutkind JS, Hellmich MR, Townsend CM: **Gastrin Stimulates Cyclooxygenase-2 Expression in Intestinal Epithelial Cells through Multiple Signaling Pathways**. *Journal of Biological Chemistry* 2002, **277**(50):48755-48763.
138. Dehez S, Daulhac L, Kowalski-Chauvel A, Fourmy D, Pradayrol L, Seva C: **Gastrin-induced DNA synthesis requires p38-MAPK activation via PKC/Ca2+ and Src-dependent mechanisms**. *FEBS letters* 2001, **496**(1):25-30.
139. Satoh A, Gukovskaya AS, Nieto JM, Cheng JH, Gukovsky I, Reeve JR, Shimosegawa T, Pandol SJ: **PKC-delta and -epsilon regulate NF-kappa B activation induced by cholecystokinin and TNF-alpha in pancreatic acinar cells**. *Am J Physiol-Gastr L* 2004, **287**(3):G582-G591.
140. Yuan JZ, Lugea A, Zheng L, Gukovsky I, Edderkaoui M, Rozengurt E, Pandol SJ: **Protein kinase D1 mediates NF-kappa B activation induced by cholecystokinin and cholinergic signaling in pancreatic acinar cells**. *Am J Physiol-Gastr L* 2008, **295**(6):G1190-G1201.
141. Koh YH, Tamizhselvi R, Bhatia M: **Extracellular Signal-Regulated Kinase 1/2 and c-Jun NH2-Terminal Kinase, through Nuclear Factor-kappa B and Activator Protein-1, Contribute to Caerulein-Induced Expression of Substance P and Neurokinin-1 Receptors in Pancreatic Acinar Cells**. *J Pharmacol Exp Ther* 2010, **332**(3):940-948.
142. Varro A, Noble P-JM, Pritchard DM, Kennedy S, Hart CA, Dimaline R, Dockray GJ: **Helicobacter pylori Induces Plasminogen Activator Inhibitor 2 in Gastric Epithelial Cells through Nuclear Factor-κB and RhoA**. *Cancer research* 2004, **64**(5):1695-1702.
143. Cordelier P, Estève JP, Bousquet C, Delesque N, O'Carroll aM, Schally aV, Vaysse N, Susini C, Buscail L: **Characterization of the antiproliferative signal mediated by the somatostatin receptor subtype sst5.** *Proceedings of the National Academy of Sciences of the United States of America* 1997, **94**:9343-9348.
144. Arena S, Pattarozzi A, Corsaro A, Schettini G, Florio T: **Somatostatin receptor subtype-dependent regulation of nitric oxide release: involvement of different intracellular pathways.** *Molecular endocrinology (Baltimore, Md)* 2005, **19**:255-267.
145. Sternfeld L, Krause E, Guse AH, Schulz I: **Hormonal control of ADP-ribosyl cyclase activity in pancreatic acinar cells from rats.** *The Journal of biological chemistry* 2003, **278**:33629-33636.
146. Lee HC: **Cyclic ADP-ribose and NAADP: fraternal twin messengers for calcium signaling.** *Science China Life sciences* 2011, **54**:699-711.

147. Fitzsimmons TJ, Gukovsky I, McRoberts Ja, Rodriguez E, Lai Fa, Pandol SJ: **Multiple isoforms of the ryanodine receptor are expressed in rat pancreatic acinar cells.** *The Biochemical journal* 2000, **351**:265-271.
148. Cancela JM, Mogami H, Tepikin aV, Petersen OH: **Intracellular glucose switches between cyclic ADP-ribose and inositol trisphosphate triggering of cytosolic Ca2+ spiking.** *Current biology : CB* 1998, **8**:865-868.
149. Sjodin L, Gardner JD: **Effect of Cholecystokinin Variant (Cck39) on Dispersed Acinar Cells from Guinea-Pig Pancreas**. *Gastroenterology* 1977, **73**(5):1015-1018.
150. Meinkoth JL, Alberts AS, Went W, Fantozzi D, Taylor SS, Hagiwara M, Montminy M, Feramisco JR: **Signal-Transduction through the Camp-Dependent Protein-Kinase**. *Mol Cell Biochem* 1993, **128**:179-186.
151. Kim C, Cheng CY, Saldanha SA, Taylor SS: **PKA-I holoenzyme structure reveals a mechanism for cAMP-dependent activation**. *Cell* 2007, **130**(6):1032-1043.
152. Naviglio S, Caraglia M, Abbruzzese A, Chiosi E, Di Gesto D, Marra M, Romano M, Sorrentino A, Sorvillo L, Spina A *et al*: **Protein kinase A as a biological target in cancer therapy**. *Expert Opin Ther Tar* 2009, **13**(1):83-92.
153. Dabrowski A, Groblewski GE, Schafer C, Guan KL, Williams JA: **Cholecystokinin and EGF activate a MAPK cascade by different mechanisms in rat pancreatic acinar cells**. *Am J Physiol-Cell Ph* 1997, **273**(5):C1472-C1479.
154. Morrison DK, Cutler Jr RE: **The complexity of Raf-1 regulation**. *Current Opinion in Cell Biology* 1997, **9**(2):174-179.
155. Zimmermann S, Moelling K: **Phosphorylation and Regulation of Raf by Akt (Protein Kinase B)**. *Science* 1999, **286**(5445):1741-1744.
156. Dabrowski A, VanderKuur JA, CarterSu C, Williams JA: **Cholecystokinin stimulates formation of Shc-Grb2 complex in rat pancreatic acinar cells through a protein kinase C-dependent mechanism**. *Journal of Biological Chemistry* 1996, **271**(43):27125-27129.
157. Daulhac L, Kowalski-Chauvel A, Pradayrol L, Vaysse N, Seva C: **Src-family tyrosine kinases in activation of ERK-1 and p85/p110-phosphatidylinositol 3-kinase by G/CCKB receptors**. *The Journal of biological chemistry* 1999, **274**(29):20657-20663.
158. Blaukat A, Ivankovic-Dikic I, Gronroos E, Dolfi F, Tokiwa G, Vuori K, Dikic I: **Adaptor proteins Grb2 and Crk couple Pyk2 with activation of specific mitogen-activated protein kinase cascades**. *Journal of Biological Chemistry* 1999, **274**(21):14893-14901.
159. Tapia JA, Ferris HA, Jensen RT, Garcia LJ: **Cholecystokinin activates PYK2/CAK beta by a phospholipase C-dependent mechanism and its association with the mitogen-activated protein kinase signaling pathway in pancreatic acinar cells**. *Journal of Biological Chemistry* 1999, **274**(44):31261-31271.

Information contained in Additional File 5

Module interactions	Details of module	All interactors	CCK model protein interactors	Ensembl Ids	HGNC symbol	Gene name	MF_GO_slim	KEGG	Reactome	
Module_NFkB Module_CCK2R Module_NFkB	IKKA_HUMAN => HMOX2_HUMAN MED29_HUMAN CBPE_HUMAN MED31_HUMAN			ENSG00000108590	MED31	mediator comp	GO:0000988_protein	NA	REACT_71:Gene Expression,	
Module_FAK1/2 module_SRC Module_EGFR r	IGRB2_HUMAN =\ FTCD_HUMAN MED29_HUMAN AIMOT_HUMAN N MED4_HUMAN			ENSG00000136146	MED4	mediator comp	GO:0008134_transcri	NA	REACT_71:Gene Expression,	
Module_NFkB Module_NFkB	IKKA_HUMAN => MED29_HUMAN MED15_HUMAN MED30_HUMAN MED7_HUMAN			ENSG00000155868	MED7	mediator comp	GO:0000988_protein	NA	REACT_71:Gene Expression,	
Module_NFkB Module_AKT1 Module_PKC mi	RHOA_HUMAN = ELOC_HUMAN MED29_HUMAN MED15_HUMAN S MED8_HUMAN			ENSG00000159479	MED8	mediator comp	GO:0000988_protein	NA	REACT_71:Gene Expression,	
Module_FAK1/2 module_SRC Module_EGFR I	IGRB2_HUMAN => RBL2_HUMAN MAGA1_HUMAN SMAD2_HUMAN SNW1_HUMAN			ENSG00000100603	SNW1	SNW domain ce	NA	hsa03040:Spl	REACT_71:Gene Expression,	
Module_AKT1 module_Rho GTPase module_	CDC42_HUMAN = KDM5A_HUMAN ATX1_HUMAN SIN3A_HUMAN Y SUH_HUMAN			ENSG00000168214	RBPJ	recombination	GO:0003677_DNA bir	hsa04330:Not	REACT_71:Gene Expression,	
Module_NFkB Module_NFkB Module_AP1 m	M3K14_HUMAN H590A_HUMAN SYK_HUMAN SYT SYDC_HUMAN			ENSG00000115866	DARS	aspartyl-tRNA t	GO:0016874_ligase a	hsa00970:Am	REACT_71:Gene Expression,	
Module_NFkB Module_AP1 module_Rho GTF	TF65_HUMAN M3K14_HUMAN TF65_HUMAN DUS YEP_HUMAN			ENSG00000136628	EPRS	glutamyl-proly	GO:0016874_ligase a	hsa00860:Por	REACT_71:Gene Expression,	
module_Rho GTPase module_BCL module_Bi	BCL2_HUMAN => CD44_HUMAN MK08_HUMAN 1433Z_HUMAN D3ISY6_HUMAN			ENSG00000106105	GARS	glycyl-tRNA syr	GO:0016874_ligase a	hsa00970:Am	REACT_71:Gene Expression,	
Module_PKC Module_PKA Module_PKA Mod	KAPCA_HUMAN M3K14_HUMAN ANDR_HUMAN 1433Z_HUMAN S IVIC_HUMAN			ENSG00000196305	IARS	isoleucyl-tRNA	GO:0016874_ligase a	hsa00290:Val	REACT_71:Gene Expression,	
Module_AKT1 module_Rho GTPase module_	CDC42_HUMAN = EF1G_HUMAN GEM4_HUMAN IKKE_HUMAN 1S51 SYK_HUMAN			ENSG00000065427	KARS	lysyl-tRNA synt	GO:0004165_ion binc	hsa00970:Am	REACT_71:Gene Expression,	
Module_PKC Module_CCK2R module_SRC M	SRC M:SRC_HUMAN => EF1G_HUMAN TF65_HUMAN 1433Z_HUMAN NEN SYLC_HUMAN			ENSG00000133706	LARS	leucyl-tRNA syr	GO:0016874_ligase a	hsa00290:Val	REACT_71:Gene Expression,	
module_SRC Module_NFkB Module_PKC Mo	KPCE_HUMAN => PPM1F_HUMAN TF65_HUMAN CSKP_HUMAN RS:SYMC_HUMAN			ENSG00000166986	MARS	methionyl-tRN	GO:0016874_ligase a	hsa00450:Sel	REACT_71:Gene Expression,	
Module_PKC Module_PKA Module_PKA Mod	KAPCA_HUMAN M3K14_HUMAN HTF4_HUMAN S SYQ_HUMAN			ENSG00000172053	QARS	glutaminyl-tRN	GO:0016874_ligase a	hsa00970:Am	REACT_71:Gene Expression,	
NA Module_FAK1/2 module_SRC Module_EG	STAT3_HUMAN = M3K14_HUMAN AIMP1_HUMAN RNPS1_HUMAN SYRC_HUMAN			ENSG00000113643	RARS	arginyl-tRNA s	GO:0016874_ligase a	hsa00970:Am	REACT_71:Gene Expression,	
Module_NFkB Module_AP1 module_Rho GTF	NFKB1_HUMAN = BCAR3_HUMAN TRAF6_HUMAN I CGI25_HUMAN Y SYTC_HUMAN			ENSG00000113407	TARS	threonyl-tRNA	GO:0016874_ligase a	hsa00970:Am	REACT_71:Gene Expression,	
Module_MAPK1/3 Module_AP1 Module_AP1	MK01_HUMAN =\ EF1G_HUMAN EF1B_HUMAN TR10D_HUMAN TRA SYVC_HUMAN			ENSG00000226589	VARS	valyl-tRNA synt	GO:0016874_ligase a	hsa00290:Val	REACT_71:Gene Expression,	
Module_PKC Module_PKA module_SRC Modi	KAPCA_HUMAN : K5YK_HUMAN ZAP70_HUMAN SHC1_HUMAN LCK PAG1_HUMAN			ENSG00000078641	PAG1	phosphoprotei	GO:0030674_protein	NA	REACT_6900:Signaling in Immune system, REACT_	
Module_AP1 Module_MAP3K11 Module_ATF	MK08_HUMAN =:Q5JY90_HUMAN TLR2_HUMAN Q69FE3_HUMAN : TIRAP_HUMAN			ENSG00000150455	TIRAP	toll-interleukin	GO:0030674_protein	hsa04620:Tol	REACT_6900:Signaling in Immune system, REACT_	
Module_CCK1R Module_AKT1 Module_CCK2R PTN11_HUMAN = ATF2_HUMAN TCAM1_HUMAN TNF10_HUMAN N IRF3_HUMAN				ENSG00000126456	IRF3	interferon regu	GO:0003677_DNA bir	hsa04620:Tol	REACT_6900:Signaling in Immune system,	
Module_FAK1/2 Module_PKC Module_CCK2R	CRK_HUMAN => IRS5_HUMAN RS10_HUMAN H2B1B_HUMAN DDX: NFKB2_HUMAN			ENSG00000077150	NFKB2	nuclear factor c	GO:0003677_DNA bir	hsa04010:MA	REACT_6900:Signaling in Immune system,	
Module_PKC Module_CCK2R module_SRC M	SRC M:SRC_HUMAN => IVRK1_HUMAN ETS2_HUMAN STRN4_HUMAN ATF:JUN_HUMAN			ENSG00000177606	JUN	jun proto-onco	GO:0030234_enzyme	hsa04010:MA	REACT_6900:Signaling in Immune system,	
Module_AP1 Module_MAP3K11 Module_ATF	MK08_HUMAN =:RS5_HUMAN IPO9_HUMAN RS10_HUMAN HCLS1 NEMO_HUMAN			ENSG00000073009	IKBKG	inhibitor of kap	GO:0043167_ion binc	hsa04010:MA	REACT_6900:Signaling in Immune system,	
Module_MAPK1/3 Module_PKC Module_PKA	PARG_HUMAN Q7GXZ8_HUMAN EFTU_HUMAN EIF3F_HUMAN B 1B07_HUMAN			ENSG00000234745	HLA-B	major histocom	GO:0004871_signal tr	hsa04144:Enc	REACT_6900:Signaling in Immune system,	
module_SRC Module_NFkB Module_PKC Mo	KPCE_HUMAN => CARD9_HUMAN TM6S4_HUMAN MALT1_HUMAN BCL10_HUMAN			ENSG00000142867	BCL10	B-cell CLL/lymp	GO:0032182_small cc	hsa04660:T cc	REACT_6900:Signaling in Immune system,	
Module_PKC Module_CCK2R module_SRC M	SRC M:SRC_HUMAN => IQSJY90_HUMAN NCK1_HUMAN GRB2_HUMAN C3WSP5_HUMAN			ENSG00000239732	TLR9	toll-like recept	NA	hsa04620:Tol	REACT_6900:Signaling in Immune system,	
NA Module_NFkB Module_AP1 module_Rho	STAT3_HUMAN =:Q5JY90_HUMAN Q59FK4_HUMAN FYN_HUMAN F CD19_HUMAN			ENSG00000177455	CD19	CD19 molecule	GO:0004871_signal tr	hsa04640:Hei	REACT_6900:Signaling in Immune system,	
Module_PKC Module_PKA module_SRC Modi	KAPCA_HUMAN : CRKL_HUMAN KPCD_HUMAN ATF5_HUMAN GREI CD34_HUMAN			ENSG00000174059	CD34	CD34 molecule	GO:0043167_ion binc	hsa04514:Cel	REACT_6900:Signaling in Immune system,	
Module_AKT1 module_Rho GTPase module_	CDC42_HUMAN = ESR1_HUMAN NCK1_HUMAN SNT CD3E_HUMAN			ENSG00000198851	CD3E	CD3e molecule	GO:0019899_enzyme	hsa04640:Hei	REACT_6900:Signaling in Immune system,	
Module_MAPK1/3 module_RAF1 module_PA	MP2K2_HUMAN : CD63_HUMAN SHC1_HUMAN TRAI CD81_HUMAN			ENSG00000110651	CD81	CD81 molecule	NA	hsa04662:B cc	REACT_6900:Signaling in Immune system,	
module_BCL module_Rho GTPase module_Bi	PAK1_HUMAN => CO5_HUMAN NCK1_HUMAN SART3_HUMAN GREC06_HUMAN			ENSG00000039537	C6	complement cc	NA	hsa04610:Cor	REACT_6900:Signaling in Immune system,	
Module_AP1 Module_MAPK1/3 Module_ATF	FOS_HUMAN => ICO5_HUMAN PLMN_HUMAN CLUS_HUMAN	CO7_HUMAN		ENSG00000112936	C7	complement cc	NA	hsa04610:Cor	REACT_6900:Signaling in Immune system,	
Module_EGFR Module_PKC Module_FAK1/2	SHC1_HUMAN => CO5_HUMAN CLUS_HUMAN CO8G_HUMAN CO8A CO8B_HUMAN			ENSG00000021852	C8B	complement cc	NA	hsa04610:Cor	REACT_6900:Signaling in Immune system,	
Module_AKT1 module_Rho GTPase module_	CDC42_HUMAN = CO8A_HUMAN CLUS_HUMAN	CO9_HUMAN		ENSG00000113600	C9	complement cc	NA	hsa04610:Cor	REACT_6900:Signaling in Immune system,	
module_RAF1 Module_PKC Module_PKA Mo	1433B_HUMAN = FCGR1_HUMAN LCK_HUMAN FCERG_HUMAN CD FCG3A_HUMAN			ENSG00000203747	FCGR3A	Fc fragment of	NA	hsa04650:Nat	REACT_6900:Signaling in Immune system,	
module_RAF1 Module_MAPK1/3 Module_PK	MP2K1_HUMAN : ERBB2_HUMAN BEND5_HUMAN LAX1_HUMAN SI GRAP2_HUMAN			ENSG00000100351	GRAP2	GRB2-related a	GO:0030674_protein	hsa04660:T cc	REACT_6900:Signaling in Immune system,	
Module_Rac-gastrin module_Rho GTPase M	CAH1_HUMAN GIT2_HUMAN PAK2Y_HUMAN KPBS1_HUMAN SLIKR_HUMAN			ENSG00000104835	NFKBIB	nuclear factor c	GO:0004871_signal tr	hsa04062:Ch	REACT_6900:Signaling in Immune system,	

Interactors / kinase / kinase regulator / phosphatase / phosphatase regulator / GTPase associated / phospholipase / adaptor / caspase regulator / protein ligase / transcripton / chromatin organizatio

Additional File 6

Table S4. The file enlists PathExpand interactors of the global CCKR pathway.

UniProt_ID	Gene name	HGNC symbol	PECompact_module
4EBP2_HUMAN	eukaryotic translation initiation factor 4E binding protein 2	EIF4EBP2	yes
4EBP3_HUMAN	eukaryotic translation initiation factor 4E binding protein 3	EIF4EBP3	yes
A8K0R3_HUMAN	osteoglycin	OGN	no
A8K5S8_HUMAN	SH2 domain containing 3C	SH2D3C	yes
AFAP1_HUMAN	actin filament associated protein 1	AFAP1	no
AKA28_HUMAN	A kinase (PRKA) anchor protein 14	AKAP14	no
ARHGP_HUMAN	Rho guanine nucleotide exchange factor (GEF) 25	ARHGEF25	yes
ASXL1_HUMAN	additional sex combs like 1 (Drosophila)	ASXL1	no
ATP5J_HUMAN	ATP synthase, H+ transporting, mitochondrial Fo complex, subunit F6	ATP5J	no
BIEA_HUMAN	biliverdin reductase A	BLVRA	no
CPSM_HUMAN	carbamoyl-phosphate synthase 1, mitochondrial	CPS1	yes
CXB1_HUMAN	gap junction protein, beta 1, 32kDa	GJB1	no
DDR2_HUMAN	discoidin domain receptor tyrosine kinase 2	DDR2	yes
DUS2_HUMAN	dual specificity phosphatase 2	DUSP2	yes
DUS22_HUMAN	dual specificity phosphatase 22	DUSP22	no
DUS4_HUMAN	dual specificity phosphatase 4	DUSP4	yes
DUS5_HUMAN	dual specificity phosphatase 5	DUSP5	yes
DUS7_HUMAN	dual specificity phosphatase 7	DUSP7	yes
DUS9_HUMAN	dual specificity phosphatase 9	DUSP9	yes
E9PDN8_HUMAN	MCF.2 cell line derived transforming sequence-like	MCF2L	yes
ERRFI_HUMAN	ERBB receptor feedback inhibitor 1	ERRFI1	yes
F263_HUMAN	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 3	PFKFB3	yes
F5GWV9_HUMAN	mucin 12, cell surface associated	MUC12	yes
FA59A_HUMAN	family with sequence similarity 59, member A	FAM59A	no
FCG2C_HUMAN	Fc fragment of IgG, low affinity IIc, receptor for (CD32) (gene/pseudogene	FCGR2C	no
GAB3_HUMAN	GRB2-associated binding protein 3	GAB3	no
GCYA3_HUMAN	guanylate cyclase 1, soluble, alpha 3	GUCY1A3	no
GNB1L_HUMAN	guanine nucleotide binding protein (G protein), beta polypeptide 1-like	GNB1L	yes
GRDN_HUMAN	coiled-coil domain containing 88A	CCDC88A	no
GRP3_HUMAN	RAS guanyl releasing protein 3 (calcium and DAG-regulated)	RASGRP3	no
GSCR1_HUMAN	glioma tumor suppressor candidate region gene 1	GLTSCR1	no
ICMT_HUMAN	isoprenylcysteine carboxyl methyltransferase	ICMT	yes
IKBZ_HUMAN	nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, zeta	NFKBIZ	no
IP3KA_HUMAN	inositol-trisphosphate 3-kinase A	ITPKA	no
K2026_HUMAN	KIAA2026	KIAA2026	no
KS6A4_HUMAN	ribosomal protein S6 kinase, 90kDa, polypeptide 4	RPS6KA4	yes
KSR1_HUMAN	kinase suppressor of ras 1	KSR1	no
MK15_HUMAN	mitogen-activated protein kinase 15	MAPK15	yes
MTPN_HUMAN	myotrophin	MTPN	yes
MUC12_HUMAN	mucin 12, cell surface associated	MUC12	yes
NCF1B_HUMAN	neutrophil cytosolic factor 1B pseudogene	NCF1B	no
NFAC3_HUMAN	nuclear factor of activated T-cells, cytoplasmic, calcineurin-dependent 3	NFATC3	no
OPHN1_HUMAN	oligophrenin 1	OPHN1	yes
P3C2B_HUMAN	phosphatidylinositol-4-phosphate 3-kinase, catalytic subunit type 2 beta	PIK3C2B	no
P4K2B_HUMAN	phosphatidylinositol 4-kinase type 2 beta	PI4K2B	yes

PAR6G_HUMAN	par-6 partitioning defective 6 homolog gamma (C. elegans)	PARD6G	no
PDE3B_HUMAN	phosphodiesterase 3B, cGMP-inhibited	PDE3B	no
PEBP4_HUMAN	phosphatidylethanolamine-binding protein 4	PEBP4	no
PHLP1_HUMAN	PH domain and leucine rich repeat protein phosphatase 1	PHLPP1	no
PKHG2_HUMAN	pleckstrin homology domain containing, family G (with RhoGef domain) member	PLEKHG2	yes
PP14A_HUMAN	protein phosphatase 1, regulatory (inhibitor) subunit 14A	PPP1R14A	no
PTPRR_HUMAN	protein tyrosine phosphatase, receptor type, R	PTPRR	no
Q0VDC6_HUMAN	FK506 binding protein 1A, 12kDa	FKBP1A	no
Q498B9_HUMAN	additional sex combs like 1 (Drosophila)	ASXL1	yes
Q53SD7_HUMAN	RAS guanyl releasing protein 3 (calcium and DAG-regulated)	RASGRP3	yes
Q59HA3_HUMAN	IQ motif containing GTPase activating protein 2	IQGAP2	yes
Q5SXQ0_HUMAN	protein tyrosine phosphatase, non-receptor type 7	PTPN7	yes
Q64GA9_HUMAN	interferon regulatory factor 5	IRF5	yes
Q6FHM9_HUMAN	CD59 molecule, complement regulatory protein	CD59	yes
Q96RR5_HUMAN	TPX2, microtubule-associated, homolog (Xenopus laevis)	TPX2	yes
Q96T11_HUMAN	receptor-interacting serine-threonine kinase 4	RIPK4	yes
Q9BTX6_HUMAN	ret proto-oncogene	RET	no
Q9NYE8_HUMAN	TGFB1-induced anti-apoptotic factor 1	TIAF1	yes
RHG31_HUMAN	Rho GTPase activating protein 31	ARHGAP31	yes
RL27_HUMAN	ribosomal protein L27	RPL27	yes
RRP5_HUMAN	programmed cell death 11	PDCD11	yes
SEC20_HUMAN	BCL2/adenovirus E1B 19kDa interacting protein 1	BNIP1	yes
SEMG2_HUMAN	semenogelin II	SEMG2	no
SH2D3_HUMAN	SH2 domain containing 3C [	SH2D3C	no
SIT1_HUMAN	signaling threshold regulating transmembrane adaptor 1	SIT1	no
STP1_HUMAN	transition protein 1 (during histone to protamine replacement)	TNP1	yes
TBC3F_HUMAN	TBC1 domain family, member 3F	TBC1D3F	yes
TRH_HUMAN	thyrotropin-releasing hormone	TRH	no
VGFR3_HUMAN	fms-related tyrosine kinase 4	FLT4	no

Information contained in Additional File 7

CCKR model interactor	CCKR model status	Modules targeted by interaction(s) by the CCKR model interactor	Number of interactions between the CCKR model interactor and the corresponding module	PathExpand prediction for the CCKR model interactor, either global or based on module	PathExpand prediction for the CCKR model interactor, either global or based on module	PathExpand prediction for the CCKR model interactor, either global or based on module	PathExpand prediction for the CCKR model interactor, either global or based on module	PathExpand prediction for the CCKR model interactor, either global or based on module
ICMT_HUMAN	NA	Modules_SRC__Defining	1	PathExpandModule__Module_AKT1	PathExpandModule__Module_PKC	PathExpandModule__Modules_Rho GTPase	PathExpandModule__Module_SRC	PathExpandGlobal
ICMT_HUMAN	NA	Modules_Rho GTPase__Incoming	1	PathExpandModule__Module_AKT1	PathExpandModule__Module_PKC	PathExpandModule__Modules_Rho GTPase	PathExpandModule__Module_SRC	PathExpandGlobal
ARHGP_HUMAN	NA	Modules_Rho GTPase__Defining	3	PathExpandModule__Module_AKT1	PathExpandModule__Module_CCK2R	PathExpandModule__Modules_Rho GTPase	PathExpandModule__Module_SRC	PathExpandGlobal
ICMT_HUMAN	NA	Modules_Rho GTPase__Defining	1	PathExpandModule__Module_AKT1	PathExpandModule__Module_PKC	PathExpandModule__Modules_Rho GTPase	PathExpandModule__Module_SRC	PathExpandGlobal
ICMT_HUMAN	NA	Modules_RAF1__Incoming	1	PathExpandModule__Module_AKT1	PathExpandModule__Module_PKC	PathExpandModule__Modules_Rho GTPase	PathExpandModule__Module_SRC	PathExpandGlobal
ARHGP_HUMAN	NA	Module_NFkB__Incoming	1	PathExpandModule__Module_AKT1	PathExpandModule__Module_CCK2R	PathExpandModule__Modules_Rho GTPase	PathExpandModule__Module_SRC	PathExpandGlobal
ICMT_HUMAN	NA	Module_NFkB__Incoming	1	PathExpandModule__Module_AKT1	PathExpandModule__Module_PKC	PathExpandModule__Modules_Rho GTPase	PathExpandModule__Module_SRC	PathExpandGlobal
ICMT_HUMAN	NA	Module_MAP3K11__Incoming	1	PathExpandModule__Module_AKT1	PathExpandModule__Module_PKC	PathExpandModule__Modules_Rho GTPase	PathExpandModule__Module_SRC	PathExpandGlobal
ARHGP_HUMAN	NA	Module_AKT1__Incoming	2	PathExpandModule__Module_AKT1	PathExpandModule__Module_CCK2R	PathExpandModule__Modules_Rho GTPase	PathExpandModule__Module_SRC	PathExpandGlobal
ICMT_HUMAN	NA	Module_AKT1__Incoming	2	PathExpandModule__Module_AKT1	PathExpandModule__Module_PKC	PathExpandModule__Modules_Rho GTPase	PathExpandModule__Module_SRC	PathExpandGlobal
RRP5_HUMAN	NA	Module_NFkB__Defining	2	PathExpandModule__Module_AP1	PathExpandModule__Module_MAP3K11	PathExpandModule__Module_NFkB	PathExpandModule__Module_PKC	PathExpandModule__Modules_Rho GTPase
F263_HUMAN	NA	Modules_SRC__Incoming	1	PathExpandModule__Module_SRC	PathExpandGlobal			
Q498B9_HUMAN	NA	Modules_SRC__Incoming	1	PathExpandModule__Module_SRC	PathExpandGlobal			

Additional File 8

Table S5: Gene Ontology molecular function terms to classify large scale protein interactors.

GO_ID	MF term description	Category
GO:0060090	binding, bridging	adaptor
GO:0030674	protein binding, bridging	adaptor
GO:0032947	protein complex scaffold	adaptor
GO:0035591	signaling adaptor activity	adaptor
GO:0005078	MAP-kinase scaffold activity	adaptor
GO:0042169	SH2 domain binding	adaptor
GO:0017124	SH3 domain binding	adaptor
GO:0030159	receptor signaling complex scaffold activity	adaptor
GO:0008093	cytoskeletal adaptor activity	adaptor
GO:0003779	actin binding	adaptor
GO:0008092	cytoskeletal protein binding	adaptor
GO:0043028	cysteine-type endopeptidase regulator activity involved in apoptotic process	caspase regulator
GO:0043274	phospholipase binding	enzyme-receptor binding
GO:0050998	nitric-oxide synthase binding	enzyme-receptor binding
GO:0002020	protease binding	enzyme-receptor binding
GO:0019902	phosphatase binding	enzyme-receptor binding
GO:0019901	protein kinase binding,	enzyme-receptor binding
GO:0005102	receptor binding	enzyme-receptor binding
GO:0030695	GTPase regulator activity	GTPase
GO:0005083	small GTPase regulator activity	GTPase
GO:0005092	GDP-dissociation inhibitor activity	GTPase
GO:0005095	GTPase inhibitor activity	GTPase
GO:0005085	guanyl-nucleotide exchange factor activity	GTPase
GO:0005096	GTPase activator activity	GTPase
GO:0003924	GTPase activity	GTPase
GO:0051020	GTPase binding	GTPase
GO:0004672	protein kinase activity	kinase
GO:0019887	protein kinase regulator activity	kinase regulator
GO:0030295	protein kinase activator activity	kinase regulator
GO:0004860	protein kinase inhibitor activity	kinase regulator
GO:0004721	phosphoprotein phosphatase activity	phosphatase
GO:0019888	protein phosphatase regulator activity	phosphatase regulator
GO:0019208	phosphatase regulator activity	phosphatase regulator
GO:0019211	phosphatase activator activity	phosphatase regulator
GO:0019212	phosphatase inhibitor activity	phosphatase regulator
GO:0004620	phospholipase activity	phospholipase
GO:0019787	small conjugating protein ligase activity	protein ligase
GO:0004842	ubiquitin-protein ligase activity	protein ligase
GO:0019788	NEDD8 ligase activity	protein ligase
GO:0019789	SUMO ligase activity	protein ligase
GO:0004888	transmembrane signaling receptor activity	receptor
GO:0005200	structural constituent of cytoskeleton	structural
GO:0005198	structural molecule activity	structural
GO:0000981	sequence-specific DNA binding RNA polymerase II transcription factor activity	transcription
GO:0003700	sequence-specific DNA binding transcription factor activity	transcription
GO:0000988	protein binding transcription factor activity	transcription
GO:0008134	transcription factor binding	transcription
GO:0003712	transcription cofactor activity	transcription
GO:0003713	transcription coactivator activity	transcription
GO:0003714	transcription corepressor activity	transcription

GO:0000989	transcription factor binding transcription factor activity	transcription
GO:0035257	nuclear hormone receptor binding	transcription
GO:0035035	histone acetyltransferase binding	transcription, chromatin organization
GO:0004407	histone deacetylase activity	transcription, chromatin organization
GO:0004402	histone acetyltransferase activity	transcription, chromatin organization
GO:0042054	histone methyltransferase activity	transcription, chromatin organization
GO:0042393	histone binding	transcription, chromatin organization
GO:0042826	histone deacetylase binding	transcription, chromatin organization
GO:0003682	chromatin binding	transcription, chromatin organization
GO:0022857	transmembrane transporter activity	transporter
GO:0008565	protein transporter activity	transporter

Paper II

BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–3

TFcheckpoint: a curated compendium of transcription factors

Konika Chawla [1] *, Sushil Tripathi [2] *, Liv Thommesen [2,3], Astrid Lægreid [2] and Martin Kuiper [1, †]

[1]Department of Biology, Norwegian University of Science and Technology (NTNU), N-7491 Trondheim, Norway.
[2]Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology (NTNU), N-7489 Trondheim, Norway.
[3]Department of Technology, Sør-Trøndelag, University College, N-7004 Trondheim, Norway.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Summary: Gene regulatory network assembly and analysis requires high quality knowledge sources that cover functional aspects of the various components of the gene regulatory machinery. A multiplicity of resources exists with information about mammalian transcription factors (TFs), yet only few of these provide sufficiently accurate classifications of the functional roles of individual TFs, or standardized evidence that would justify the information on which these functional classifications are based. We compiled the list of all putative TFs from 9 different resources, and checked available literature for references that support their function as a true sequence-specific DNA-binding TF (DbTF). The results are available in the TFcheckpoint database, an exhaustive collection of TFs annotated according to experimental and other evidence on their function as true DbTFs. TFcheckpoint.org provides a high quality and comprehensive knowledge source for genome scale regulatory network studies.

Availability: The TFcheckpoint database is freely available at www.tfcheckpoint.org

Contact: martin.kuiper@ntnu.no

Supplementary material: Supplementary information is available at Bioinformatics online.

1 INTRODUCTION

Transcription factors (TFs) lie at the basis of gene-expression diversity in different cell-types and conditions. TFs constitute key gene regulatory components that usually participate in large multiprotein-DNA complexes, where they guide RNA Polymerase (i.e. RNAP I, II and III) activity and regulate the onset and rate of RNA synthesis. These protein complexes may include general transcription factors (GTFs) that bind to core-promoter DNA; general co-factors that bind to GTFs to form a pre-initiation transcription complex; specific DNA-binding transcription factors and factors that lack DNA-binding domains but exert their regulatory roles through interaction with other proteins in the transcription complex. This last class of protein-interacting transcription regulators includes co-activators, co-repressors, histone modifiers and chromatin remodeling proteins (Lee and Young, 2000).

The DNA-binding transcription factors (DbTFs) play a central role in specifying which genes are transcribed, as they guide the transcription machinery to distinct target genes by binding to specific gene regulatory elements located in proximal promoters as well as in distal enhancer regions (Mitchell and Tjian, 1989). The DbTF proteins that regulate RNA Polymerase II (RNAP II) enjoy a special focus in gene regulatory network building due to their strong ability to explain the protein coding gene expression landscape of biological responses. Access to accurate and genome-scale knowledge concerning these DbTFs therefore is of key importance. Multiple resources with knowledge about mammalian transcription factors exist (Harris *et al.*, 2004; Fulton *et al.*, 2009; Kummerfeld and Teichmann, 2006; Messina *et al.*, 2004; Ravasi *et al.*, 2010; Sandelin *et al.*, 2004; Schaefer *et al.*, 2011; Vaquerizas *et al.*, 2009; Zhang *et al.*, 2012), however, we observed that 1) most of them do not distinguish well between true DbTFs, protein-interacting TFs and general TFs and 2) only in a minority of cases do they provide standardized evidence for the functional role of the TFs. Because of this, users of these resources will only have an obscured view at the domain of DbTFs. Here we present TFcheckpoint (www.tfcheckpoint.org), a comprehensive repository of human, mouse and rat TF candidates. All entries have been manually checked for literature information pertaining to their potential biological function as DbTFs. The database serves as a checkpoint for TF information, is freely available and supports ID or name searching, browsing and bulk download.

2 RESULTS

2.1 Database content

TFcheckpoint contains the cumulative inventory of 9 major TF information sources (Fig. 1 and Supplementary Material), and we manually checked each of these entries for literature describing evidence for RNAP II-regulating DNA binding transcription factors,

*authors contributed equally

†to whom correspondence should be addressed

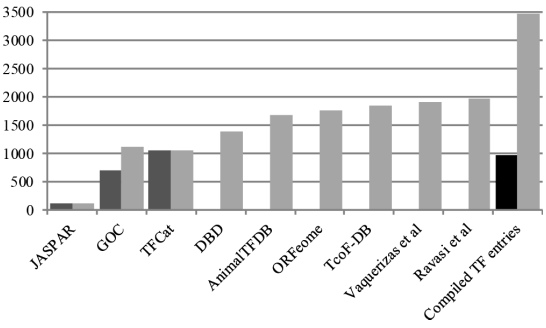


Fig. 1. TF candidate and associated literature references. For each resource, the total numbers of TF entries (light grey) and TF entries with literature references (dark grey) are given. For GOC data, all unique proteins annotated to "Sequence specific DNA-binding transcription factor activity (GO:0003700)" or any of its children, are listed. The bar to the far right indicates 3462 unique TF entries in TFcheckpoint; 983 of these (adjacent black bar) were deemed to be true DbTFs.

for human, mouse or rat. The evidence that we selected should at least support Gene Ontology term "Sequence-specific DNA-binding RNA polymerase II transcription factor activity (GO:0000981)", taking this as the minimum defining term for a true RNAPII regulating DbTF. In general we selected the first PubMed paper(s) that showed satisfactory evidence for a specific TF (for details see Supplementary Material).

We assembled a list of 3462 putative TFs from the above resources (Fig 1). We have used orthology mappings from UniProt to identify corresponding gene Entrez IDs from human, rat and mouse. For 983 proteins we could identify one or more relevant papers with the evidence for a DbTF, yielding a total of 1072 unique PubMed references. 824 DbTFs are supported by literature references with experimental evidence, whereas a further 154 are supported by author statements and a final 5 are supported by sequence based analysis. The full list and the literature reference results are available from the TFcheckpoint database.

The availability of high quality and exhaustive information at one central place facilitates the access by the global scientific community. We are currently working together with the Gene Ontology Consortium to develop and apply general standards for TF annotation and merge our findings with the GO database (Harris *et al.*, 2004).

2.2 Database user interface

TFcheckpoint is powered by MySQL and accessible through a web interface created with Joomla (http://www.joomla.org), implementing HTML and PHP scripts. The database is hosted on an apache server at the Norwegian data infrastructure Norstore (http://www.norstore.no), and available at www.tfcheckpoint.org. The database can be used for simple browsing of all 3462 candidate TFs as well as the subset of DbTFs with literature evidence. For each DbTF the literature reference(s) and information about the original TF candidate resource that we obtained it from are provided. All TF entries are linked to Entrez and UniProt IDs. The NCBI official gene symbol is used as a primary key, but the data can also be searched for any of the NCBI provided synonyms, as well as Entrez and UniProt identifiers. All data is also downloadable as a tab-delimited text file.

3 CONCLUSION

A literature-based exhaustively curated list of transcription factors is an invaluable resource for researchers working on gene regulatory mechanisms. The ENCODE project is targeting the generation of evidenice for some 3000 putative TFs (ENCODE Project Consortium *et al.*, 2012). Our current list of curated TFs provides a reference both for small scale experiments and genome-scale studies where researchers either need to verify predicted lists of TFs even before characterizing the role of these regulatory proteins (Choi *et al.*, 2006; Gray *et al.*, 2004) or use background knowledge of TFs to infer gene regulatory networks (Ye *et al.*, 2009). By ensuring that these annotations become part of the GO database this knowledge will become available to all analysis approaches based on Gene Ontology knowledge.

ACKNOWLEDGEMENT

This work was funded by the Norwegian University of Science and Technology, Trondheim, Norway.

Conflicts of Interest: : none declared.

REFERENCES

Choi, M. Y., Romer, A. I., Hu, M., *et al.* (2006). A dynamic expression survey identifies transcription factors relevant in mouse digestive tract development. *Development*, **133**(20), 4119–29.

ENCODE Project Consortium, Dunham, I., Kundaje, A., *et al.* (2012). An integrated encyclopedia of dna elements in the human genome. *Nature*, **489**(7414), 57–74.

Fulton, D. L., Sundararajan, S., Badis, G., *et al.* (2009). Tfcat: the curated catalog of mouse and human transcription factors. *Genome Biol*, **10**(3), R29.

Gray, P. A., Fu, H., Luo, P., *et al.* (2004). Mouse brain organization revealed through direct genome-scale tf expression analysis. *Science*, **306**(5705), 2255–7.

Harris, M. A., Clark, J., Ireland, A., *et al.* (2004). The gene ontology (go) database and informatics resource. *Nucleic Acids Res*, **32**(Database issue), D258–61.

Kummerfeld, S. K. and Teichmann, S. A. (2006). Dbd: a transcription factor prediction database. *Nucleic Acids Res*, **34**(Database issue), D74–81.

Lee, T. I. and Young, R. A. (2000). Transcription of eukaryotic protein-coding genes. *Annu Rev Genet*, **34**, 77–137.

Messina, D. N., Glasscock, J., Gish, W., and Lovett, M. (2004). An orfeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression. *Genome Res*, **14**(10B), 2041–7.

Mitchell, P. J. and Tjian, R. (1989). Transcriptional regulation in mammalian cells by sequence-specific dna binding proteins. *Science*, **245**(4916), 371–8.

Ravasi, T., Suzuki, H., Cannistraci, C. V., *et al.* (2010). An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, **140**(5), 744–52.

Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W., and Lenhard, B. (2004). Jaspar: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*, **32**(Database issue), D91–4.

Schaefer, U., Schmeier, S., and Bajic, V. B. (2011). Tcof-db: dragon database for human transcription co-factors and transcription factor interacting proteins. *Nucleic Acids Res*, **39**(Database issue), D106–10.

Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A., and Luscombe, N. M. (2009). A census of human transcription factors: function, expression and evolution. *Nat Rev Genet*, **10**(4), 252–63.

Ye, C., Galbraith, S. J., Liao, J. C., and Eskin, E. (2009). Using network component analysis to dissect regulatory networks mediated by transcription factors in yeast. *PLoS Comput Biol*, **5**(3), e1000311.

Zhang, H.-M., Chen, H., Liu, W., *et al.* (2012). Animaltfdb: a comprehensive animal transcription factor database. *Nucleic Acids Res*, **40**(Database issue), D144–9.

Supplementary material to: TFcheckpoint: a curated compendium of transcription factors

Supplementary Material

Introduction:

In this document we provide additional information about the different sources that we checked to obtain the input list of candidate transcription factors (TFs) for the TFcheckpoint database. In addition, we provide some details about the text curation effort that we undertook to establish the existence of experimental evidence that would support a DNA-binding transcription factor (DbTF) classification of the putative TFs.

a) Overview of TF-data sources

We built a cumulative list of putative TFs by collecting all entries marked as transcription factor from sources with mammalian TFs (see Supplementary Table 1). This provided a list of 3462 unique putative TFs that we next checked for functional evidence in literature.

Supplementary Table 1. Transcription factor sources.

Sources	Entries	Species	URL	PubMed / Version / Date
TFCat (Fulton, et al., 2009)	1052	human, mouse, rat	http://www.tfcat.ca/	PMID: 19284633/ Release 1.0 / March 12, 2009
JASPAR (Sandelin, et al., 2004)	115	human, mouse, rat	http://jaspar.cgb.ki.se/	PMID: 18006571 / October 12, 2009
DBD (Kummerfeld and Teichmann, 2006)	1395	human, mouse, rat	http://www.transcriptionfactor.org/index.cgi?Home	PMID: 16381970 / Release 2.0
ORFeome (Messina, et al., 2004)	1770	human		PMID: 15489324 / October, 2004
AnimalTFDB (Zhang, et al., 2011)	1681	human, mouse, rat	http://115.156.249.50/TFDB/index.php	PMID: 22080564 / November 12, 2011
Vaquerizas et al (Vaquerizas, et al., 2009)	1909	human		PMID: 19274049 / April, 2009
Ravasi et al (Ravasi, et al., 2010)	1967	human, mouse		PMID: 20211142 / March 05, 2010
TcoF-DB (Schaefer, et al., 2011)	1860	human	http://cbrc.kaust.edu.sa/tcof/index.php	PMID: 20965969 / October 2010
GOC (Harris, et al., 2004)	1120	Human, mouse, rat	http://amigo.geneontology.org	PMID: 10802651 / February 16, 2013

Data from 9 sources (column 1) were downloaded and used to assemble a comprehensive list of proposed TFs. The table shows the identifier(s) of the source; the number of unique entries obtained from that source, the species, the URL if the source is a database, the PubMed ID of the appropriate reference and the time of download.

Supplementary material to: TFcheckpoint: a curated compendium of transcription factors

b) DbTF annotation procedure

A DbTF by definition binds to specific DNA sequences and regulates the transcription of the gene that it binds to. Therefore, in our DbTF annotation procedure we considered the following two functional properties as the minimum criteria to qualify a protein as DbTF:

i) there is evidence that the protein binds to specific DNA sequences and

ii) the protein has been demonstrated to be involved in RNAPII dependent regulation of transcription.

Next, we compiled a list of experimental assays for protein-DNA interaction and transcription regulation in order to identify the above evidence types for TFs in scientific publications.

Then we looked for specific scientific publications that would contain evidence to qualify TFs according to our DbTF annotation criteria. We started checking the already existing TF annotations by inspecting the literature that their annotations referred to. The majority of these existing annotations came from GOC (174 DbTFs), JASPAR (112 DbTFs) and TFCat (231 DbTFs). Next, we searched the literature for experimental evidence supporting the remaining TF candidates, by performing searches for gene names in the following resources: UniProt (http://www.uniprot.org/), NCBI's Entrez Gene (Maglott, et al., 2007), iHOP (Hoffmann and Valencia, 2004), Gene Cards (Safran, et al., 2002), and NCBI's PubMed (http://www.ncbi.nlm.nih.gov/pubmed/). This yielded additional literature references for 466 DbTFs.

References

Fulton, D.L., *et al.* (2009) TFCat: the curated catalog of mouse and human transcription factors, *Genome biology*, **10**, R29.

Harris, M.A., *et al.* (2004) The Gene Ontology (GO) database and informatics resource, *Nucleic Acids Research*, **32**, D258-D261.

Hoffmann, R. and Valencia, A. (2004) A gene network for navigating the literature, *Nature genetics*, **36**, 664.

Kummerfeld, S.K. and Teichmann, S.A. (2006) DBD: a transcription factor prediction database, *Nucleic Acids Research*, **34**, D74-D81.

Maglott, D., *et al.* (2007) Entrez Gene: gene-centered information at NCBI, *Nucleic Acids Res*, **35**, D26-31.

Messina, D.N., *et al.* (2004) An ORFeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression, *Genome Res*, **14**, 2041-2047.

Ravasi, T., *et al.* (2010) An atlas of combinatorial transcriptional regulation in mouse and man, *Cell*, **140**, 744-752.

Safran, M., *et al.* (2002) GeneCards 2002: towards a complete, object-oriented, human gene compendium, *Bioinformatics*, **18**, 1542-1543.

Sandelin, A., *et al.* (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles, *Nucleic Acids Res*, **32**, D91-94.

Schaefer, U., Schmeier, S. and Bajic, V.B. (2011) TcoF-DB: dragon database for human transcription co-factors and transcription factor interacting proteins, *Nucleic Acids Research*, **39**, D106-D110.

Vaquerizas, J.M., *et al.* (2009) A census of human transcription factors: function, expression and evolution, *Nature reviews. Genetics*, **10**, 252-263.

Zhang, H.-M., *et al.* (2011) AnimalTFDB: a comprehensive animal transcription factor database, *Nucleic Acids Research*.

Paper III

Gene Ontology Annotation of Sequence specific DNA-binding Transcription Factors: Setting the Stage for a Large Scale Curation Effort

Sushil Tripathi[1], Karen R. Christie[2], Rama Balakrishnan[3], Rachael Huntley[4], David P. Hill[2], Liv Thommesen[1,5], Judith A. Blake[2], Martin Kuiper[6], Astrid Lægreid[1]*

[1] Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology (NTNU), N-7489 Trondheim, Norway.

[2] Department of Mammalian Genetics, The Jackson Laboratory, 600 Main Street, Bar Harbor, ME, USA.

[3] Department of Genetics, Stanford University, Stanford, CA, 94305-5120, USA.

[4] EMBL-EBI, The Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK.

[5] Department of Technology, Sør-Trøndelag University College, N-7004 Trondheim, Norway.

[6] Department of Biology, Norwegian University of Science and Technology (NTNU), N-7491 Trondheim, Norway.

*Corresponding author

Institute of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology (NTNU), N-7489 Trondheim, Norway.

Tel.: +4772825323

Fax : +4772571463;

E-mail : astrid.lagreid@ntnu.no

Abstract

Transcription factors control which information in a genome becomes transcribed in order to produce RNAs that function in the biological systems of cells and organisms. Reliable and comprehensive information about transcription factors is invaluable for large-scale network based studies. However, existing transcription factor knowledge bases are still lacking in well documented functional information.

Here we provide guidelines for a curation strategy, which constitutes a robust framework for using the controlled vocabularies defined by the Gene Ontology Consortium (GOC) to annotate specific DNA binding transcription factors based on experimental evidence reported in literature. Our standardized protocol and workflow for annotating specific DNA binding RNA polymerase II transcription factors (DbTFs) is designed to document high quality and decisive evidence from valid experimental methods. Within a collaborative biocuration effort involving the user community, we are now in the process of exhaustively annotating the full repertoire of human, mouse and rat proteins that qualify as DbTFs in as much as they are experimentally documented in the biomedical literature today. The completion of this task will significantly enrich Gene Ontology based information resources for the research community.

Introduction

Specific gene regulation mechanisms determine which part of the genome becomes transcribed in order to provide the active molecular parts of living organisms in various environmental conditions. Central in these mechanisms are multiprotein complexes present at the regulatory regions of genes that determine onset and rate of RNA synthesis by regulating RNA polymerase activity (1, 2). These multiprotein complexes comprise general transcription factors (GTFs), general co-factors (3), RNA polymerase II (RNAP II) sequence-specific DNA binding transcription factors (DbTFs) (4), and a large array of transcription factors that lack DNA-binding activity but exert their regulatory roles through protein interaction with the aforementioned proteins and that include co-activators, co-repressors, histone modifiers and chromatin remodeling proteins (1, 2). GTFs bind to core-promoter DNA where they constitute pre-initiation transcription complexes (PICs), in concert with general co-factors, whereas DbTFs bind to gene-specific proximal and distal gene regulatory regions. RNAP II, one of the three nuclear RNA polymerases (RNAP I, II and III) involved in transcription

of mammalian genes, draws special attention in studies directed at gene regulatory mechanisms since it is responsible for transcribing protein-coding genes as well as miRNA genes (5).

Due to their selective binding within regulatory regions of distinct genes, the DbTFs play decisive roles in directing the assembly of the multiprotein transcription machinery to a particular subset of genes. This assembly can either be followed by immediate RNAP II dependent transcription or it can result in promoter-proximal pausing of RNAP II that can subsequently be released into active transcription triggered by either DbTFs or by other mechanisms (6, 1, 7). DbTFs also play a central role in transcription repression either by competing with activating DbTFs for DNA binding or by recruiting transcriptional co-repressors (8, 2). Through these functions, DbTFs link the phenotypical state of the cell - reflected in abundance and activation state of proteins in the transcriptional machinery - to the decoding of regulatory information embedded within the genome sequence. Thus, the DbTFs are a point of convergence for mechanisms involved in upward causation, i.e. the flow of information from genome to phenome (central dogma), as well as in downward causation, which enables the organism to respond to cues from the extrinsic and intrinsic environment (9).

Current estimates suggest that the human genome contains ~1900 DbTF-coding genes (10). With the increasing trend to pursue systems-level understanding of gene regulatory networks (11) it is of key importance to have available genome-wide and accurate information concerning DbTFs including their specific roles in transcription regulation, their target genes (TGs) and their expression patterns related to cell type and to developmental as well as to normal- and pathophysiological processes. This need for genome-wide information has sparked among others the ENCODE project, an initiative to identify all functional elements in the human genome sequence as well as regulatory interactions between TFs and their transcription factor binding sites (TFBS) (12). Thus, experimental data will continue to become available in ever increasing volumes, and subsequent comprehensive annotation of functional aspects of DbTFs in public databases will be of high value for ongoing and future gene regulatory studies.

Gene Ontology (GO) provides a common vocabulary for the functional description of genes and gene products and consists of three sub-ontologies: Biological Process (BP), Molecular Function (MF) and Cellular Component (CC) (13). The Gene Ontology Consortium (GOC) provides high quality classifications for types of transcription factors and captures the supporting evidence for the assignment of classes to gene products. Recently (2010-2011) the GOC undertook a major reorganization of the representation of transcription factors within GO to bring this area up to date with current knowledge. This process generated a more accurate ontology structure utilizing newly introduced definitions to define precise relationships between terms (14). For example, since nucleic acid-binding transcription factors must bind nucleic acid as part of their function, Molecular Function (MF) terms for types of *"nucleic acid binding transcription factor activity"* have "has_part" relationships to the appropriate MF terms for *"nucleic acid binding"* (e.g. *"sequence-specific DNA binding RNA polymerase II transcription factor activity" (GO:0000981)* has_part *"RNA polymerase II regulatory region sequence-specific DNA binding" (GO:0000977)*).Similarly, MF *"transcription factor activity"* terms (e.g. *"sequence-specific DNA binding RNA polymerase II transcription factor activity" (GO:0000981)*) have "part_of" relationships to appropriate Biological Process (BP) terms for *"regulation of transcription"* (e.g. *"regulation of transcription from RNA polymerase II promoter" (GO:0006357)*), since another required aspect of a DNA-binding transcription factor lies in its role in regulating transcription. These "part_of" relationships are shown in Figure 1.

Today[1], GOC provides annotations that allow for identification of ~300 human, mouse, and rat DbTFs which is about 15% of the expected DbTFs (10). A mere ~200 of these are presently supported by experimental evidence, while ~100 are annotated with evidence based on computational prediction, sequence and structure similarity or author statement[1]. There are several mammalian DbTF databases, including TFcat (15), JASPAR (16), and TFe (17) that also hold experimentally documented DbTF-information based on cited scientific literature. But the information in these databases lacks the informative annotations founded on ontologies

[1] GO database release on 16th Feb. 2013

and evidence codes as provided by the GOC and necessary for rigorous computational reasoning and analysis.

The above findings suggest that to date no single comprehensive information resource for mammalian DbTFs exists with the level of coverage and high-quality annotation that is needed for genome-scale data analysis and interpretation. The GOC has standard procedures for annotating proteins, and their database is authoritative in providing comprehensive annotations to the myriad of tools that use GO information for data analysis. However, the capacity of expert curators at the GOC is presently not scaled for or focused on dedicated efforts to comprehensively annotate one particular functional protein class. Therefore, we have embarked on a collaborative effort involving community users and GOC members to exhaustively curate experimentally documented mammalian DbTFs. Similar to other sub-domain annotation initiatives (18, 19), our first aim was to develop specific guidelines for curating experimentally documented DbTFs from literature. This included the assembly of a list of experimental assays that would qualify to provide verifiable functional evidence for genuine DbTFs. Here, we provide a detailed report in the form of a comprehensive curation protocol, based on which we currently are engaged in a focused effort to curate all DbTFs from a collection of candidate proteins compiled from the major TF information sources. A database providing detailed information about TF information sources and assembled DbTF documentation is available at www.tfcheckpoint.org.

Creation of annotations for sequence specific DNA binding RNAPII Transcription Factors (DbTFs)

Our curation guidelines for high quality annotation of experimentally verified DbTFs are designed to capture the essential functional capabilities of DbTFs and record published evidence using rigorous semantics. In the following sections we describe fundamental functional characteristics of a DbTF, how these characteristics can adequately be described by Gene Ontology terms, and how these terms and evidence codes can be asserted based on experimental work reported in literature. The assembled procedure facilitates a precise representation of DbTF functional attributes using the standard GOC defined gene-association file format (GAF2.0;

http://www.geneontology.org/GO.format.gaf-2_0.shtml) and the PSI-MI data exchange format used for recording interaction data (20). A detailed DbTF annotation guideline document is provided in Supplementary material.

Criteria that qualify a DbTF

A DbTF is a DNA binding transcription factor that binds to a specific DNA sequence and regulates the transcription of the associated gene. The specific DNA sequences bound by DbTFs are termed transcription factor binding sites (TFBS) and are located in gene regulatory regions either upstream and proximal to the core promoter, or in more distal upstream or downstream enhancer regions. Once a DbTF recognizes a TFBS it may recruit other accessory factors or RNAPII or it may interfere with binding of other regulatory proteins to regulate the expression of the target gene. This means that a DbTF must exhibit both DNA-binding and transcription regulation capacity. Therefore, the minimum criteria to qualify a protein as DbTF for RNAPII are that it: i) binds to specific DNA sequences in gene regulatory regions and ii) is involved in RNAPII-dependent regulation of transcription. It is evident that in order to capture these functional aspects accurately and efficiently, the specific Gene Ontology terms that substantiate these assertions need to be precisely defined. These GO terms must address both "sequence specific DNA binding" and "transcription regulation" capabilities accurately. In the following sections, we provide a detailed reasoning behind the selection of specific GO terms of different granularity as well as assignment of GO evidence codes and experimental assays that are considered adequate and necessary for creating a DbTF annotation.

Gene Ontology terms used for DbTF annotation

Specific DNA binding

To capture the capability of a protein to bind to specific DNA sequences, GO molecular function (MF) terms are used that describe *"sequence-specific DNA-binding"* such as *GO:0043565 (sequence-specific DNA binding)*, and *GO:0000977 (RNA polymerase II regulatory region sequence-specific DNA binding)* of which the latter depicts binding to any portion of the regulatory sequence for a gene transcribed by RNA polymerase II. Whenever information is available indicating whether the protein binds proximal or distal regulatory regions,

the terms *GO:0000978 (RNA polymerase II core promoter proximal region sequence-specific DNA binding)* or *GO:0000980 (RNA polymerase II distal enhancer sequence-specific DNA binding)* are used (see Figure 1 A, terms colored yellow).

Involvement in RNAPII dependent regulation of transcription

The involvement of a protein in transcription regulation is well captured by the GO biological process (BP) terms *GO:0006357 (regulation of transcription from RNA polymerase II promoter)* or any of its children that specify whether the protein is involved in positive or negative regulation of transcription (see Figure 1A, terms in blue).

Sequence specific DNA binding RNAP II transcription factor activity

The goal of this curation project is to assign a DNA binding transcription factor activity term, i.e. *GO: 0000981 (sequence-specific DNA binding RNA polymerase II transcription factor activity)* or one of its children to appropriate DbTFs (Figure 1A, terms colored green). As indicated above, this requires that the composite functional aspects of DbTF proteins: DNA binding and transcription regulation; must each be represented by their proper MF and BP GO terms. These different aspects of DbTF activity: DNA binding and involvement in transcriptional regulation are typically demonstrated in different experiments, sometimes not even presented in the same paper, so the annotations to DNA binding (MF) and transcriptional regulation (BP) terms are made separately, and only when both are assigned (each in their inherent logic of the GO-structure) can they be combined to infer DbTF activity molecular function terms (Table 1).

The child terms of *GO: 0000981* are used to delineate whether the TF exerts its activity by binding to the promoter proximal region or the distal enhancer , i.e. *GO:0000982 (RNA polymerase II core promoter proximal region sequence-specific DNA binding transcription factor activity)* or *GO:0003705 (sequence-specific distal enhancer binding RNA polymerase II transcription factor activity)* and whether the result of binding is positive or negative regulation of target gene transcription e.g. *GO:0001077 (RNA polymerase II core promoter proximal region sequence-specific DNA binding transcription factor activity involved in positive regulation of transcription)* and *GO:0001205 (RNA polymerase II distal enhancer sequence-specific*

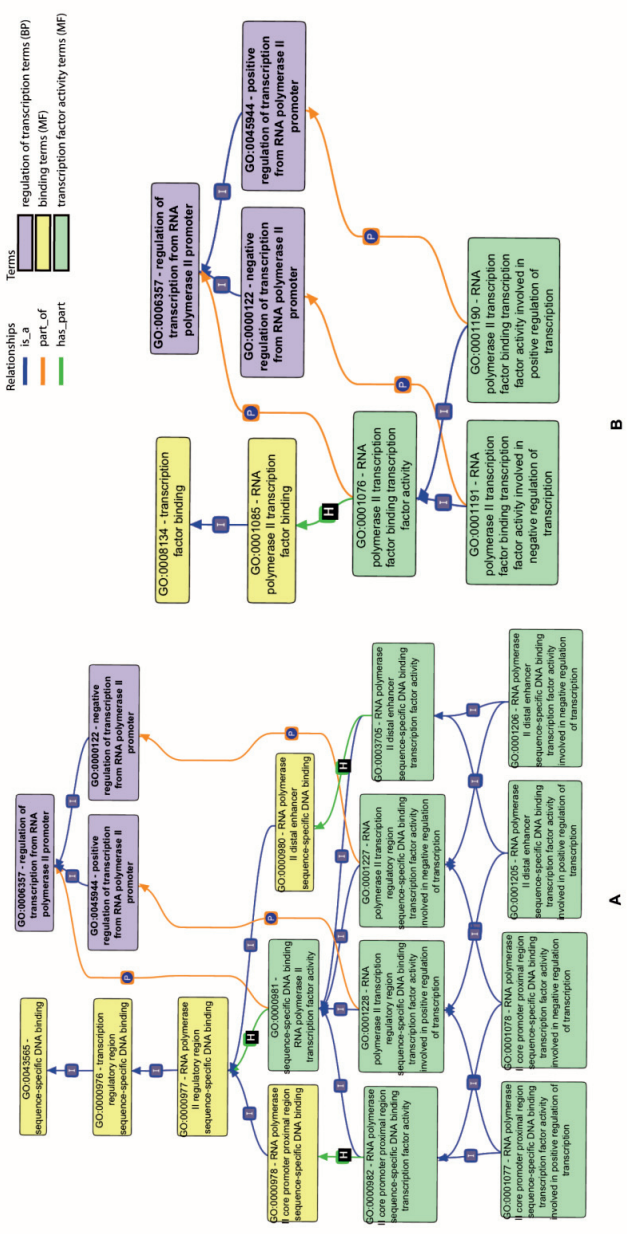


Figure 1: Primary GO terms/subgraphs used for DbTF annotation. (A) GO subgraph used for sequence-specific DNA binding transcription factors. **(B)** GO subgraph used for transcription factor binding transcription factors.

DNA binding transcription factor activity involved in positive regulation of transcription)

Table 1: Inference of DbTF activity terms from DNA binding- and transcription regulation terms. Each transcription factor activity term (green) is determined by the composite annotation of corresponding DNA binding term (yellow) and transcription regulation terms (blue).

DNA binding terms (MF)	***Transcription regulation terms (BP)***		
	GO:0006357 regulation of transcription from RNA polymerase II promoter	***GO:0045944 positive regulation of transcription from RNA polymerase II promoter***	***GO:0000122 negative regulation of transcription from RNA polymerase II promoter***
GO:00043565 sequence-specific DNA binding	**GO: 0000981** sequence-specific DNA binding RNA polymerase II transcription factor activity	**GO:0001228** RNA polymerase II transcription regulatory region sequence-specific DNA binding transcription factor activity involved in positive regulation of transcription	**GO:0001227** RNA polymerase II transcription regulatory region sequence-specific DNA binding transcription factor activity involved in negative regulation of transcription
GO:0000977 RNA polymerase II regulatory region sequence-specific DNA binding	**GO: 0000981** sequence-specific DNA binding RNA polymerase II transcription factor activity	**GO:0001228** RNA polymerase II transcription regulatory region sequence-specific DNA binding transcription factor activity involved in positive regulation of transcription	**GO:0001227** RNA polymerase II transcription regulatory region sequence-specific DNA binding transcription factor activity involved in negative regulation of transcription
GO:0000978 RNA polymerase II core promoter proximal region sequence-specific DNA binding	**GO:0000982** RNA polymerase II core promoter proximal region sequence-specific DNA binding transcription factor activity	**GO:0001077** RNA polymerase II core promoter proximal region sequence-specific DNA binding transcription factor activity involved in positive regulation of transcription	**GO:0001078** RNA polymerase II core promoter proximal region sequence-specific DNA binding transcription factor activity involved in negative regulation of transcription
GO:0000980 RNA polymerase II distal enhancer sequence-specific DNA binding	**GO:0003705** sequence-specific distal enhancer binding RNA polymerase II transcription factor activity	**GO:0001205** RNA polymerase II distal enhancer sequence-specific DNA binding transcription factor activity involved in positive regulation of transcription	**GO:0001206** RNA polymerase II distal enhancer sequence-specific DNA binding transcription factor activity involved in negative regulation of transcription

TF-binding and TF-binding TF activity

Transcriptional regulation mechanisms are complex. Usually many TFs work together in concert to regulate transcription. In instances where the activity of a TF is reported to be dependent on interaction with another protein or multi-subunit complex, the protein-protein interaction is annotated using *"transcription factor binding"* molecular function GO terms as shown in Figure 1B (terms in yellow). Furthermore, a different set of *"transcription factor activity"* terms, i.e. *GO:0001076 (RNA polymerase II transcription factor binding transcription factor activity)* or any of its children, is chosen reflecting the fact that the activity is dependent on binding to another TF (Figure 1B, terms with green color).

Once TF-binding and transcription regulation are each annotated individually, the GO structure allows generating TF-binding *TF* activity annotations by combining the separate annotations (Table 2).

Table 2. Inference of TF binding activity terms from TF binding and transcription regulation. Each TF-binding transcription factor activity term (green) is determined by the composite annotation of corresponding TF binding term (yellow) and transcription regulation term (blue).

TF binding terms (MF)	***Transcription regulation terms (BP)***		
	GO:0006357 regulation of transcription from RNA polymerase II promoter	***GO:0045944 positive regulation of transcription from RNA polymerase II promoter***	***GO:0000122 negative regulation of transcription from RNA polymerase II promoter***
GO: 0008134 Transcription factor binding	**GO: 0001076** RNA polymerase II transcription factor binding transcription factor activity	**GO:0001190** RNA polymerase II transcription factor binding transcription factor activity involved in positive regulation of transcription	**GO:0001191** RNA polymerase II transcription factor binding transcription factor activity involved in negative regulation of transcription
GO: 0001085 RNA polymerase II transcription factor binding	**GO: 0001076** RNA polymerase II transcription factor binding transcription factor activity	**GO:0001190** RNA polymerase II transcription factor binding transcription factor activity involved in positive regulation of transcription	**GO:0001191** RNA polymerase II transcription factor binding transcription factor activity involved in negative regulation of transcription

When the functional unit of a TF is a complex

In instances where the complex is a homodimer, or higher order multimer of the same protein, there are no special annotation issues as all of the activities demonstrated are properties of the same gene product. However, when the functional unit is a heterodimer or other multi-subunit complex, then there are some additional considerations for annotation.

The "contributes to" qualifier is specifically intended for the annotation of functions that occur in the context of complexes, rather than being an activity of a single subunit within the complex. In the case of a heterodimer, there are times where one of the two proteins does not bind DNA on its own. However, in some cases a subunit that does not bind DNA independently can be shown to contribute to the sequence specificity of binding when present within a heterodimer. In this situation, the subunit that does not bind DNA alone could be annotated to appropriate "sequence-specific DNA binding" terms (Figure 1A, in yellow) using the qualifier "contributes to" to indicate that it contributes to the DNA binding of the heterodimer. More generally, the "contributes to" qualifier can be used in conjunction with any MF term, including the "transcription factor activity" terms, to indicate that it contributes to that function within the context of a complex, even though it does not possess that activity independently. In contrast, in a multi-subunit TF where the DNA binding activity is known to be confined to one or more specific subunits, other subunits should not be annotated to a "DNA binding" term at all.

For any subunit within a TF complex, it is appropriate to annotate all appropriate GO terms for which that function has been experimentally shown, either individually or as part of the complex indicated with the "contributes to" qualifier. Thus, in some cases, a given protein may be annotated both with a *"sequence specific DNA binding RNAP II transcription factor activity"* term as well as with a 'TF binding RNAP II transcription factor activity' term.

Evidence codes and experimental assays

In accordance with the overall guidelines for GO annotations, each DbTF annotation must be qualified with an evidence code indicating how the annotation is supported by experimental evidence (http://www.geneontology.org/GO.evidence.shtml). The DbTF curation guidelines presented in the current work use one of the following GO evidence codes: Inferred from Direct Assay (IDA), Inferred from Physical Interaction (IPI), Inferred from Mutant Phenotype (IMP), or Inferred by Curator (IC).

When a single scientific paper comprises all experimental evidence necessary to support each of the annotations for 'DNA- or TF-binding' and 'Transcription regulation', the evidence codes for these two annotations are transferred to the composite DbTF annotation to a MF *"transcription factor activity"* term (see Table 3). However, when the two different types of annotations ('DNA - or 'TF-binding' and 'transcription regulation') for a given TF cannot be generated from one single paper, the evidence code 'Inferred by Curator' (IC) is used along with the GOC generated reference, GO_REF:0000036 (http://www.geneontology.org/cgi-bin/references.cgi#GO_REF:0000036). The IC code, which requires use of the two GO IDs for the appropriate 'binding' and 'transcription regulation' terms, indicates that GO-annotations based on evidence from two different sources have been combined by a curator to infer the appropriate transcription factor activity term.

Table 3: Evidence code table. Transcription factor activity evidence code is selected based on the evidence for DNA-binding/TF-binding (MF) term and transcription regulation (BP) term.

DNA binding/-TF-binding	Transcription regulation	TF activity
IDA	IDA	IDA/IC*
IMP	IMP	IMP/IC*
IDA	IMP	IDA, IMP/IC*
IMP	IDA	IMP, IDA/IC*
IPI**	IDA	IPI**, IDA/ IC*

IC* if evidence for 'DNA binding / TF-binding' and 'transcription regulation' comes from two different papers

IPI** applicable only for TF-binding terms

In order to provide for a uniform standard for evaluation of experimental evidence for DbTF annotations we surveyed several relevant resources defining experimental assays that can document TF function, including ORegAnno (21), TRRD (22), RegulonDB (23), and the PSI-MI controlled vocabulary for molecular interactions (20).

In the following sections we have compiled selected sets for the experimental assays that we deemed to be most relevant for annotation of DNA binding, TF binding and transcription regulation. PSI-MI-unique identifiers are given wherever they exist. Augmentation of the PSI-MI vocabulary to span a larger repertoire of TF-defining experiments is ongoing.

Specific DNA-binding

Experimental data documenting specific DNA binding is obtained from experiments that show *in vitro* binding of a TF to specific DNA sequences present in either cloned TG regulatory regions (proximal promoter and/or distal enhancer) or in synthetic DNA sequences representing canonical TF binding sites or specific TG regulatory regions (see Table 4). We have chosen not to rely on assays measuring *in vivo* TF-DNA interaction (e.g. the ChIP (Chromatin ImmunoPrecipitation) assay) because it is not possible to ascertain in these assays that the TF in question actually binds directly to DNA, or whether some other component in the *in vivo* system mediates the TF-DNA-association.

The *in vitro* assay that has been most frequently used for documenting sequence-specific binding of TF is the Electrophoretic Mobility Shift Assay (EMSA) (24). The most common variants of this assay present the TF in the form of

i) nuclear extract from native tissue or cells
ii) nuclear extracts from cells or tissue with ectopic expression of a TF
iii) purified TF (*in vitro* translated or purified from cell extract)
iv) nuclear extract from cells with ectopic expression of a mutated TF
v) purified mutated TF (*in vitro* translated or purified from cell extract)

When the TF is presented in any of the variants ii – v, the EMSA qualifies for annotation of a GO term for ‘specific DNA binding’. In the case where the TF is presented as a nuclear extract from native cells or tissue (i), we require that the specific TF is identified with an additional

experimental approach. This may involve the use of a TF-specific antibody (EMSA supershift), or specific competition experiments demonstrating that the EMSA gel shift is not abolished by competition with an unlabeled DNA probe with a point mutation in a known TFBS for this specific TF, whereas competition with unlabeled DNA probe containing the wild type TFBS does abolish the gel shift. If no additional experimental verification of the TF is reported, nuclear extract based EMSAs of type i) do not suffice to qualify DNA binding properties of a TF, and the experiment needs to be dismissed. Similarly, the other assays listed in Table 4 must have been performed in a manner that provides for identification of the specific TF tested and to assess specific interaction between this TF and a specified DNA probe. For MI:0114 X-ray crystallography, to qualify as experimental evidence of a TFs DNA binding, it is required that the protein is co-crystallized with a DNA sequence that represents either a canonical TFBS or an authentic gene regulatory region.

Table 4: Assays documenting specific DNA binding.

Experimental assays	Evidence code	PSI-MI code
Electrophoretic mobility shift assay (EMSA)	IDA	MI:0413
Electrophoretic mobility supershift assay (EMSA supershift)	IDA	MI:0412
Footprinting	IDA	MI:0417
DNase I footprinting (DNA footprint)	IDA	MI:0606
Methylation interference assay (MIC)	IDA	MI:1189
Ultraviolet (uv) footprinting (UV-footprint)	IDA	MI:1191
Dimethylsulphate footprinting (DMS-footprint)	IDA	MI:0603
Hydroxy radical footprinting (Hydroxy-footprint)	IDA	MI:1190
Potassium permanganate footprinting (KMnO4-footprint)	IDA	MI:0604
Affinity chromatography technology	IDA	MI:0004
Pull down	IDA	MI:0096
Southwestern blot assay (SW-blot)	IDA	
In vitro evolution of nucleic acids (SELEX)	IDA	MI:0657
X-ray crystallography	IDA	MI:0114

RNAP II transcription regulation

The 'transcription regulation' terms need support from assays that document modulation of transcriptional process in response to TF action. These assays mainly fall into two groups: either reporter gene assays measuring the transcriptional regulatory effect of a TF on a regulatory region cloned upstream of a reporter gene (for instance luciferase, beta-galactosidase, or chloramphenicol acetyltransferase (CAT)), or measurement of expression levels of a target gene mRNA (see Table 5). Within each of the assays a variety of experimental strategies can allow for the identification of the specific TF ('e.g. knock in' (ectopic expression) and/or 'knock down'). Furthermore, the gene regulatory region can be presented and assessed in different ways in the reporter gene assays (e.g. 'canonical TFBS' or 'authentic TG promoter/enhancer') and different methods used to assay mRNA expression levels of specific TGs. The combinations of different modes of TF and TG detection together define the GO evidence codes to be used (Table 5).

Table 5. Reporter gene-based assays variants documenting transcription regulation. This table a decision matrix for selecting GO evidence codes based on the method used for TF identification (purple) and transcription regulation (green).

TF identification	Transcription regulation assays						
	reporter gene assay				*TG expression assay*		
	canonical TFBS	*authentic TG promoter*	*authentic TG promoter with TFBS point mutation*	*authentic TG promoter with deletion mutations*	*primer specific PCR*	*northern blot*	*ribonuclease assay*
wt TF overexpression	IDA	IDA	IDA	IDA	IDA	IDA	IDA
mut TF overexpression	IMP	IMP	IMP	IMP	IMP	IMP	IMP
TF knock down (RNAi/antisense RNA)	IMP	IMP	IMP	IMP	IMP	IMP	IMP

Whereas the experimental assays depicted in Table 5 are most often carried out by transfecting expression and reporter plasmids into cell line model systems, transcription regulation annotations can also be supported by whole organism experiments, e.g. knock out

mutations or RNAi knock down strategies. However, as such experiments do not by themselves prove a role in regulation of transcription; such annotations must be made with caution and will depend on a strict awareness of additional information such as e.g. the concomitant documentation of specific binding by the protein in question, to regulatory regions of an RNAP II regulated gene.

TF-binding

'TF-binding' specific terms are based on any assay that provides evidence for protein-protein interactions. Table 6 lists experimental assays and evidence codes that are eligible for TF-binding specific terms.

Table 6: Assays documenting TF binding.

Assays	evidence Code	PSI-MI code
2-hybrid interactions	IPI	MI:0018
Co-purification	IPI, IDA	MI:0004
Co-immunoprecipitation	IPI, IDA	MI:0019

Annotating target genes (TGs)

An obvious important biological property of a TF lies in the particular TGs that it regulates. Proper recording of this information is of key importance for the building of gene regulatory networks. In studies of DbTF functionality, often one or several specific target genes will be identified and experimentally documented. The Gene Ontology Consortium has introduced an Annotation Extension field to capture additional information that provides more biological context to the GO annotation (GAF 2.0, http://www.geneontology.org/GO.format.gaf-2_0.shtml). This field can be used to record information regarding specific TGs regulated by the TF that is being annotated. The TG is recorded in the Annotation Extension field for the BP transcription regulation GO term using the *"has_regulation_target"* relationship combined with the gene identifier(s) for the target gene(s).

Work flow of annotation

The annotation workflow is depicted in Figure 2. An annotation effort typically starts with one of the scientific papers suggested in databases such as TFCat and JASPAR to document a candidate DbTF, or by searching for adequate literature in one of the following resources: UniProt (http://www.uniprot.org/), NCBI's Entrez Gene (25), iHOP (26), Gene Cards (27), or NCBI's PubMed (http://www.ncbi.nlm.nih.gov/pubmed/).

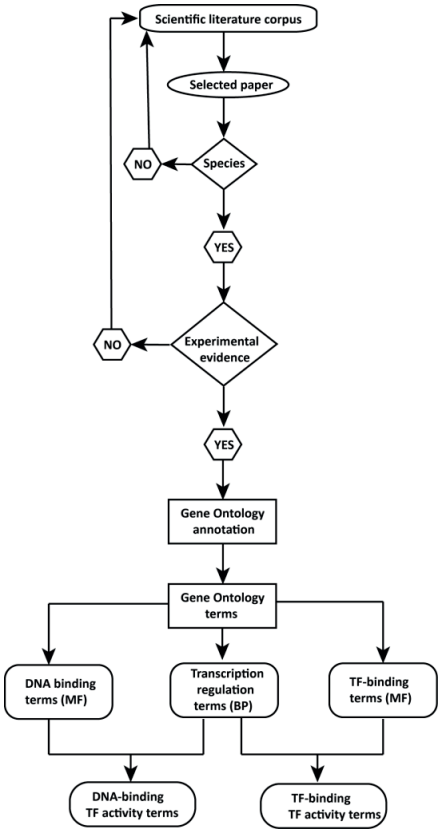
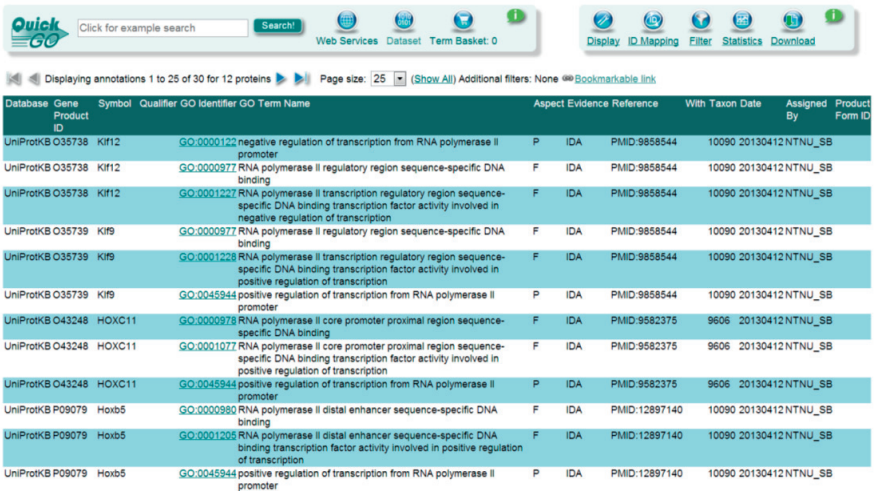


Figure 2: Sequence - specific DNA binding TF (DbTF) curation workflow. This workflow represents the step-by-step procedure for curating experimentally verified mammalian DbTFs from scientific publications. Selection of scientific publication from the literature corpus is the starting point of the curation procedure. From each relevant publication, DbTF specific GO-terms are annotated and recorded.

Each scientific paper is first checked for information providing correct identification of species origin of the TF studied. Since we are focusing on DbTFs from human, mouse, and rat studies, only papers allowing identification of a DbTF from one of these species will proceed to further curation. Then, adequate experimental evidence to support one or several DbTF annotations is searched. If either TF species origin or sufficient experimental evidence is not identifiable, the curator returns to the scientific literature corpus to search for other suitable papers. When both criteria are fulfilled, the individual GO annotations (i.e. DNA-binding and/or TF-binding and transcription regulation) are assigned together with a supporting evidence code. Finally, the composite TF activity GO terms is inferred. TF annotation data are submitted to UniProt-GOA in the form of a gene association file (GAF2.0; http://www.geneontology.org/GO.format.gaf-2_0.shtml) and will subsequently appear in the GOC database (Figure 3).



Database	Gene Product ID	Symbol	Qualifier	GO Identifier	GO Term Name	Aspect	Evidence	Reference	With	Taxon	Date	Assigned By	Product Form ID
UniProtKB	O35738	Klf12		GO:0000122	negative regulation of transcription from RNA polymerase II promoter	P	IDA	PMID:9858544		10090	20130412	NTNU_SB	
UniProtKB	O35738	Klf12		GO:0000977	RNA polymerase II regulatory region sequence-specific DNA binding	F	IDA	PMID:9858544		10090	20130412	NTNU_SB	
UniProtKB	O35738	Klf12		GO:0001227	RNA polymerase II transcription regulatory region sequence-specific DNA binding transcription factor activity involved in negative regulation of transcription	F	IDA	PMID:9858544		10090	20130412	NTNU_SB	
UniProtKB	O35739	Klf9		GO:0000977	RNA polymerase II regulatory region sequence-specific DNA binding	F	IDA	PMID:9858544		10090	20130412	NTNU_SB	
UniProtKB	O35739	Klf9		GO:0001228	RNA polymerase II transcription regulatory region sequence-specific DNA binding transcription factor activity involved in positive regulation of transcription	F	IDA	PMID:9858544		10090	20130412	NTNU_SB	
UniProtKB	O35739	Klf9		GO:0045944	positive regulation of transcription from RNA polymerase II promoter	P	IDA	PMID:9858544		10090	20130412	NTNU_SB	
UniProtKB	O43248	HOXC11		GO:0000978	RNA polymerase II core promoter proximal region sequence-specific DNA binding	F	IDA	PMID:9582375		9606	20130412	NTNU_SB	
UniProtKB	O43248	HOXC11		GO:0001077	RNA polymerase II core promoter proximal region sequence-specific DNA binding transcription factor activity involved in positive regulation of transcription	F	IDA	PMID:9582375		9606	20130412	NTNU_SB	
UniProtKB	O43248	HOXC11		GO:0045944	positive regulation of transcription from RNA polymerase II promoter	P	IDA	PMID:9582375		9606	20130412	NTNU_SB	
UniProtKB	P09079	Hoxb5		GO:0000980	RNA polymerase II distal enhancer sequence-specific DNA binding	F	IDA	PMID:12897140		10090	20130412	NTNU_SB	
UniProtKB	P09079	Hoxb5		GO:0001205	RNA polymerase II distal enhancer sequence-specific DNA binding transcription factor activity involved in positive regulation of transcription	F	IDA	PMID:12897140		10090	20130412	NTNU_SB	
UniProtKB	P09079	Hoxb5		GO:0045944	positive regulation of transcription from RNA polymerase II promoter	P	IDA	PMID:12897140		10090	20130412	NTNU_SB	

Figure 3: UniProt-GOA screenshot of some of the DbTF annotations generated using the DbTF curation guidelines discussed here.

Discussion

Benefits of a focused annotation project

A comprehensive resource of high quality annotations of TFs is of high value both for small-scale experiments where it is important to select an optimal subset of relevant TFs as well as for genome-scale studies. In the latter case, access to extensive background knowledge for TFs is essential to infer gene regulatory networks (28) or to design experiments to characterize this group of proteins as a functional class in a system-wide approach (29, 30).

Compilation and in-depth analysis of available information on transcription factors indicates that more than 800 mammalian DbTFs are experimentally documented in the scientific literature (www.tfcheckpoint.org). The current work aims to provide the foundation to curate this source of information and to record adequate GO annotations in compliance with the standards defined here. Currently[2] only 202 human, mouse, and rat proteins are annotated as DbTFs with *GO:0000981* (or any of its child terms) supported by experimental evidence; meaning that some 600 DbTFs still need to be processed. We aim to complete this task before the end of 2013. Even though the number of curators involved is small, the efficiency of this focused annotation project is high, since the number of different GO terms and evidence codes is limited and well defined, thus allowing each curator to process a relatively high number of scientific papers (typically 5 papers or more per working day).

Added value of rigorous classification of experimental assay requirements for the annotations

The catalogue of experimental assays that qualify for supporting TF annotations presented here is assembled based on the extensive TF annotation experience in the collaborating organizations. This aspect of the annotation procedure improves the quality of the GO annotations since it provides a uniform standard for interpretation of evidence strength in published experimental work. As some of the assays presently are not adequately covered by PSI-MI vocabulary (20), part of our efforts have been directed to collaborate with the PSI-MI consortium to develop additional PSI-MI terms.

[2] GO database release on 16th Feb. 2013

The proper documentation of experimental evidence for each TF annotation will enable us to work towards submitting annotated data to the IntAct database (31). Moreover, we plan to make the experimental assay details for the TF annotations available to users via our TF database (http://www.tfcheckpoint.org/). This will enable users to select subsets of TFs based on the specific experimental methods used to characterize them.

Concluding remarks

The fact that of formalized knowledge representation metadata are rarely presented in biomedical publications often makes it difficult for a curator to extract accurate information for ontology- or structured vocabulary-annotation from natural language used in the literature. The GOC provides not only guidelines for the curation of gene products information from scientific publications, but also procedures for identification of the type of evidence that supports the curated information. Because of these standardized conventions, literature-curated data in the GO database is deemed to be of high quality. In the present work, we have established a comprehensive and, specific curation procedure for TFs of RNA polymerase II which, similar to other data standardization initiatives, provides details on the requirements to properly record an experimentally verified DbTF.

The GOC is centrally involved in efforts to provide annotation guidelines for particular protein functional categories. However, the elaboration of procedures for specific tasks like the curation of distinct functional categories of proteins, or of biological process subdomains, is enhanced when experts in the respective fields are involved in the curation process. Moreover, the active participation from domain experts is greatly facilitated by generating detailed curation guidelines as vehicles for productive interactions. With the transcription factor curation effort presented here we wish to provide not only a greater number of high quality annotations for DbTFs and their TGs across three mammalian species, but also to exemplify the constructive use of detailed guidelines to facilitate collaborative biocuration efforts across institutions.

Funding

This work was supported by The Norwegian Cancer Society, The Liaison Committee between the Central Norway Regional Health Authority (RHA), the Norwegian University of Science and Technology (NTNU) and Sør-Trøndelag University College (HisT).

Conflict of interest

None declared.

References

1. Weake,V.M. and Workman,J.L. (2010) Inducible gene expression: diverse regulatory mechanisms. *Nature reviews. Genetics*, **11**, 426–37.

2. Perissi,V., Jepsen,K., Glass,C.K. and Rosenfeld,M.G. (2010) Deconstructing repression: evolving models of co-repressor action. *Nature reviews. Genetics*, **11**, 109–23.

3. Thomas,M.C. and Chiang,C.-M. (2006) The general transcription machinery and general cofactors. *Critical reviews in biochemistry and molecular biology*, **41**, 105–78.

4. Mitchell,P.J. and Tjian,R. (1989) Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science (New York, N.Y.)*, **245**, 371–8.

5. Lee,Y., Kim,M., Han,J., Yeom,K.-H., Lee,S., Baek,S.H. and Kim,V.N. (2004) MicroRNA genes are transcribed by RNA polymerase II. *The EMBO journal*, **23**, 4051–60.

6. Rahl,P.B., Lin,C.Y., Seila,A.C., Flynn,R. a, McCuine,S., Burge,C.B., Sharp,P. a and Young,R. a (2010) c-Myc regulates transcriptional pause release. *Cell*, **141**, 432–45.

7. Adelman,K. and Lis,J.T. (2012) Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nature reviews. Genetics*, **13**, 720–31.

8. Thiel,G., Lietz,M. and Hohl,M. (2004) How mammalian transcriptional repressors work. *European journal of biochemistry / FEBS*, **271**, 2855–62.

9. Noble,D. (2012) A theory of biological relativity: no privileged level of causation. *Interface focus*, **2**, 55–64.

10. Vaquerizas,J.M., Kummerfeld,S.K., Teichmann,S. a and Luscombe,N.M. (2009) A census of human transcription factors: function, expression and evolution. *Nature reviews. Genetics*, **10**, 252–63.

11. Walhout,A.J.M. (2006) Unraveling transcription regulatory networks by protein-DNA and protein-protein interaction mapping. *Genome research*, **16**, 1445–54.

12. Neph,S., Vierstra,J., Stergachis,A.B., Reynolds,A.P., Haugen,E., Vernot,B., Thurman,R.E., John,S., Sandstrom,R., Johnson,A.K., et al. (2012) An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, **489**, 83–90.

13. Gene Ontology Consortium (2013) Gene Ontology annotations and resources. *Nucleic acids research*, **41**, D530–5.

14. Leonelli,S., Diehl,A.D., Christie,K.R., Harris,M.A. and Lomax,J. (2011) How the gene ontology evolves. *BMC bioinformatics*, **12**, 325.

15. Fulton,D.L., Sundararajan,S., Badis,G., Hughes,T.R., Wasserman,W.W., Roach,J.C. and Sladek,R. (2009) TFCat: the curated catalog of mouse and human transcription factors. *Genome biology*, **10**, R29.

16. Sandelin,A., Alkema,W., Engström,P., Wasserman,W.W. and Lenhard,B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic acids research*, **32**, D91–4.

17. Yusuf,D., Butland,S.L., Swanson,M.I., Bolotin,E., Ticoll,A., Cheung,W. a, Zhang,X.Y.C., Dickman,C.T.D., Fulton,D.L., Lim,J.S., et al. (2012) The transcription factor encyclopedia. *Genome biology*, **13**, R24.

18. Inglis,D.O., Skrzypek,M.S., Arnaud,M.B., Binkley,J., Shah,P., Wymore,F. and Sherlock,G. (2013) Improved gene ontology annotation for biofilm formation, filamentous growth, and phenotypic switching in Candida albicans. *Eukaryotic cell*, **12**, 101–8.

19. Mutowo-Meullenet,P., Huntley,R.P., Dimmer,E.C., Alam-Faruque,Y., Sawford,T., Jesus Martin,M., O'Donovan,C. and Apweiler,R. (2013) Use of Gene Ontology Annotation to understand the peroxisome proteome in humans. *Database : the journal of biological databases and curation*, **2013**, bas062.

20. Hermjakob,H., Montecchi-Palazzi,L., Bader,G., Wojcik,J., Salwinski,L., Ceol,A., Moore,S., Orchard,S., Sarkans,U., Von Mering,C., et al. (2004) The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data. *Nature biotechnology*, **22**, 177–83.

21. Griffith,O.L., Montgomery,S.B., Bernier,B., Chu,B., Kasaian,K., Aerts,S., Mahony,S., Sleumer,M.C., Bilenky,M., Haeussler,M., et al. (2008) ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic acids research*, **36**, D107–13.

22. Kolchanov,N. a, Ignatieva,E. V, Ananko,E. a, Podkolodnaya,O. a, Stepanenko,I.L., Merkulova,T.I., Pozdnyakov,M. a, Podkolodny,N.L., Naumochkin, a N. and Romashchenko, a G. (2002) Transcription Regulatory Regions Database (TRRD): its status in 2002. *Nucleic acids research*, **30**, 312–7.

23. Gama-Castro,S., Jiménez-Jacinto,V., Peralta-Gil,M., Santos-Zavaleta,A., Peñaloza-Spinola,M.I., Contreras-Moreira,B., Segura-Salazar,J., Muñiz-Rascado,L., Martínez-Flores,I., Salgado,H., et al. (2008) RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic acids research*, **36**, D120–4.

24. Woo,A.J., Dods,J.S., Susanto,E., Ulgiati,D. and Abraham,L.J. (2002) A proteomics approach for the identification of DNA binding activities observed in the electrophoretic mobility shift assay. *Molecular & cellular proteomics : MCP*, **1**, 472–8.

25. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2011) Entrez Gene: gene-centered information at NCBI. *Nucleic acids research*, **39**, D52–7.

26. Hoffmann,R. and Valencia,A. (2004) CORRESPONDENCE A gene network for navigating the literature To the editor : **36**, 65102.

27. Safran,M., Solomon,I., Shmueli,O., Lapidot,M., Shen-Orr,S., Adato,A., Ben-Dor,U., Esterman,N., Rosen,N., Peter,I., et al. (2002) GeneCards 2002: towards a complete, object-oriented, human gene compendium. *Bioinformatics (Oxford, England)*, **18**, 1542–3.

28. Ye,C., Galbraith,S.J., Liao,J.C. and Eskin,E. (2009) Using network component analysis to dissect regulatory networks mediated by transcription factors in yeast. *PLoS computational biology*, **5**, e1000311.

29. Choi,M.Y., Romer,A.I., Hu,M., Lepourcelet,M., Mechoor,A., Yesilaltay,A., Krieger,M., Gray,P. a and Shivdasani,R. a (2006) A dynamic expression survey identifies transcription factors relevant in mouse digestive tract development. *Development (Cambridge, England)*, **133**, 4119–29.

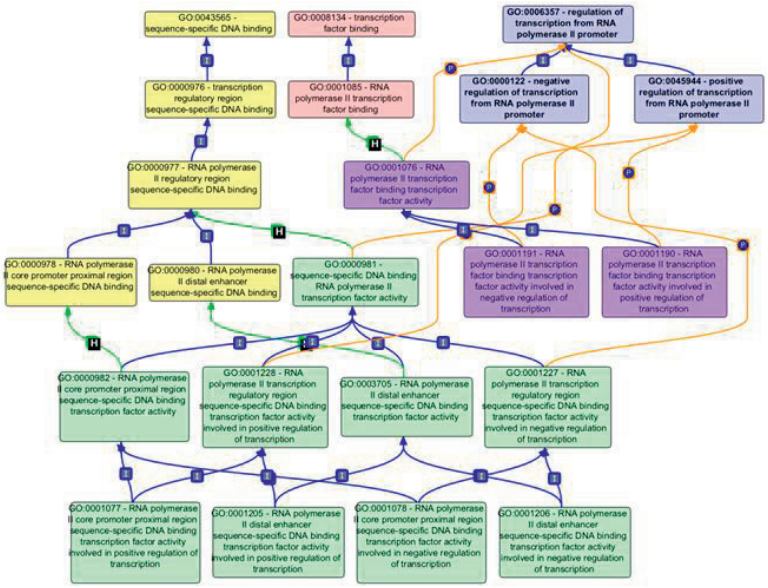
30. Gray,P. a, Fu,H., Luo,P., Zhao,Q., Yu,J., Ferrari,A., Tenzen,T., Yuk,D.-I., Tsung,E.F., Cai,Z., et al. (2004) Mouse brain organization revealed through direct genome-scale TF expression analysis. *Science (New York, N.Y.)*, **306**, 2255–7.

31. Kerrien,S., Aranda,B., Breuza,L., Bridge,A., Broackes-Carter,F., Chen,C., Duesbury,M., Dumousseau,M., Feuermann,M., Hinz,U., et al. (2012) The IntAct molecular interaction database in 2012. *Nucleic acids research*, **40**, D841–6.

Curation Guidelines

- **Sequence-specific DNA binding RNA polymerase II transcription factors**
- **Transcription factor binding RNA polymerase II transcription factors**

Content ***Page***



Curation workflow

Identify TF species

The current curation guidelines are focused on the mammalian species: human, mouse, rat

The species that the TF (i.e. its coding sequence) originates from must be unequivocally determined. If the publication used for curation does not state this explicitly, TF references must be traced for species determination or authors contacted to obtain relevant information.
If it is not possible to assert the species of the TF studied, then the paper cannot be used for curation

Create annotations for DNA binding, Transcription Regulation, TF binding

Identify an experiment that qualifies for annotation. Use Assay look-up Tables (pages 2 - 6) to assess eligibility and GO evidence code.

(optional) Assign experimental assay term and PSI-MI code

Use Assay look-up Tables (pages 2 - 6) to assign experimental assay, its variant and PSI-MI code

Create annotations for DbTF activity or TF binding activity

Create these terms by combining the DNA (or TF) binding and transcription regulation terms generated in the steps above.

Use '*Decision Tables'* for DNA binding TFs (Table 3, page 4) or TF binding TFs (Table 6, page 6)

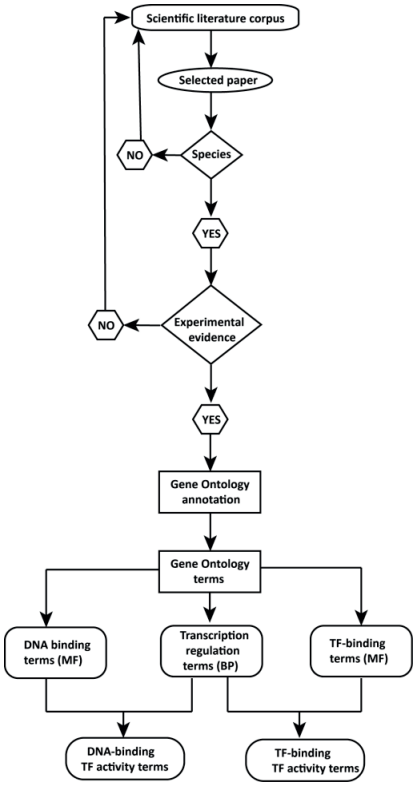


Figure 1: Curation workflow

DNA binding annotations

Use Table 1 to identify valid experimental evidence for one of the GO terms for DNA-binding shown in Figure 2

Whenever possible, choose *GO:0000977 - RNA polymerase II regulatory region sequence-specific DNA binding* or one of its children. Use *GO:00043565* only when it is not possible to identify information stating that the specific DNA sequence bound by the protein is found in a gene regulatory region, and *GO:0000976* only when it is not possible to identify information stating that the regulatory region containing the DNA-sequence specifically bound by the protein is not part of a gene regulated by RNA polymerase II.

DNA binding detection methods differ in how the TF is presented (as detailed for EMSA variants in Table 1). Presentation of TF as a nuclear extract from native cells or tissue is not sufficient for annotation of DNA binding *("No evid.").* In these instances we search for other experimental evidence that can identify the specific TF, e.g. TF-specific antibody in EMSA supershift.

X-ray crystallography used as evidence for DNA binding requires that the TF is co-crystallized with a DNA sequence that represents either a canonical TFBS or an authentic gene regulatory region.

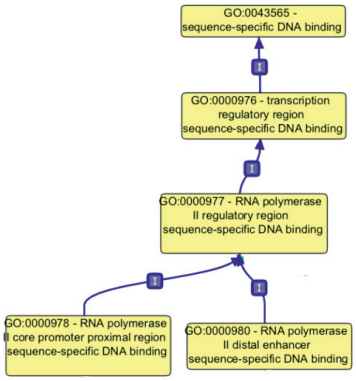


Figure 2: GO terms for MF DNA binding

Table 1: Assays and evidence codes for Molecular Function DNA binding terms

Experimental assays	Variants	Evidence code	PSI-MI code
Electrophoretic mobility shift assay (EMSA)	nuclear extract from native tissue or cells	*No evid.*	MI:0413
	nuclear extracts from cells or tissue with ectopic expression of a TF	IDA	MI:0413
	purified TF (*in vitro* translated or purified from cell extract)	IDA	MI:0413
	nuclear extract from cells with ectopic expression of a mutated TF	IMP	MI:0413
	purified mutated TF (*in vitro* translated or purified from cell extract	IMP	MI:0413
Electrophoretic mobility supershift assay (EMSA supershift)	nuclear extract from native tissue or cells	IDA	MI:0412
	nuclear extracts from cells or tissue with ectopic expression of a TF	IDA	MI:0412
	purified TF (*in vitro* translated or purified from cell extract)	IDA	MI:0412
	nuclear extract from cells with ectopic expression of a mutated TF	IMP	MI:0412
	purified mutated TF (*in vitro* translated or purified from cell extract)	IMP	MI:0412
Footprinting		IDA	MI:0417
DNase I footprinting (DNA footprint)		IDA	MI:0606
Methylation interference assay (MIC)		IDA	MI:1189
Ultraviolet (uv) footprinting (UV-footprint)		IDA	MI:1191
Dimethylsulphate footprinting (DMS-footprint)		IDA	MI:0603
Hydroxy radical footprinting (Hydroxy-footprint)		IDA	MI:1190
Potassium permanganate footprinting (KMnO4-footprint)		IDA	MI:0604
Affinity chromatography technology		IDA	MI:0004
Pull down		IDA	MI:0096
Southwestern blot assay (SW-blot)		IDA	
In vitro evolution of nucleic acids (SELEX)		IDA	MI:0657
X-ray crystallography		IDA	MI:0114

Transcription Regulation annotations

Use Table 2 to identify valid experimental evidence for one of the GO terms for transcription regulation shown in Figure 3.

Whenever possible, indicate whether the regulation is positive or negative, by using the adequate GO terms

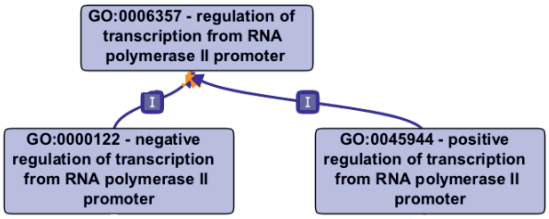


Figure 3: GO terms for BP transcription regulation

Transcription regulation annotations can also be supported by **whole organism experiments**, e.g. knock out mutations or RNAi knock down strategies.

Such experiments do not by themselves prove a role in regulation of transcription and must therefore be made with caution: they depend on a strict awareness of additional information such as the concomitant documentation of specific binding by the protein in question, to regulatory regions of an RNAP II regulated gene.

Table 2: Assays and evidence codes for Biological Process Transcription regulation terms

	Transcription regulation assays						
	reporter gene assay				***TG expression assay***		
TF identification	*canonical TFBS*	*authentic TG promoter*	*authentic TG promoter with TFBS point mutation*	*authentic TG promoter with deletion mutations*	*primer specific PCR*	*northern blot*	*ribo-nuclease assay*
					MI:0088	MI:0929	MI:0920
wt TF overexpression	IDA	IDA	IDA	IDA	IDA	IDA	IDA
mut TF overexpression	IMP	IMP	IMP	IMP	IMP	IMP	IMP
TF knock down (RNAi/antisense RNA)	IMP	IMP	IMP	IMP	IMP	IMP	IMP

DNA binding TF activity annotations

Based on the protein's existing GO annotations for specific DNA-binding (MF) and for transcription regulation (BP), create a DNA-binding TF activity (MF) annotation.

Use Decision Table 3 to identify the correct GO term (shown in green in Figure 4).

Assign evidence codes according to Table 4 – see next page.

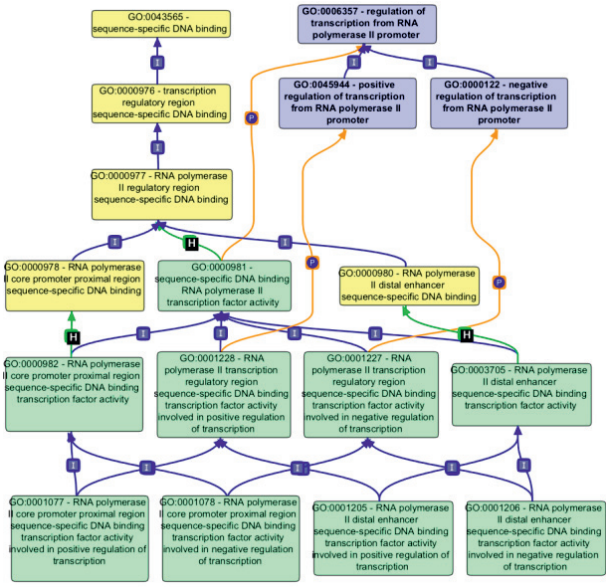


Figure 4: GO terms for MF DNA binding TF activity

Table 3: Decision table DNA binding TF activity

DNA binding terms (MF)	*Transcription regulation terms (BP)*		
	GO:0006357 regulation of transcription from RNA polymerase II promoter	***GO:0045944 positive regulation of transcription from RNA polymerase II promoter***	***GO:0000122 negative regulation of transcription from RNA polymerase II promoter***
GO:00043565 sequence-specific DNA binding	GO: 0000981 sequence-specific DNA binding RNA polymerase II transcription factor activity	GO:0001228 RNA polymerase II transcription regulatory region sequence-specific DNA binding transcription factor activity involved in positive regulation of transcription	GO:0001227 RNA polymerase II transcription regulatory region sequence-specific DNA binding transcription factor activity involved in negative regulation of transcription
GO:0000977 RNA polymerase II regulatory region sequence-specific DNA binding	GO: 0000981 sequence-specific DNA binding RNA polymerase II transcription factor activity	GO:0001228 RNA polymerase II transcription regulatory region sequence-specific DNA binding transcription factor activity involved in positive regulation of transcription	GO:0001227 RNA polymerase II transcription regulatory region sequence-specific DNA binding transcription factor activity involved in negative regulation of transcription
GO:0000978 RNA polymerase II core promoter proximal region sequence-specific DNA binding	GO:0000982 RNA polymerase II core promoter proximal region sequence-specific DNA binding transcription factor activity	GO:0001077 RNA polymerase II core promoter proximal region sequence-specific DNA binding transcription factor activity involved in positive regulation of transcription	GO:0001078 RNA polymerase II core promoter proximal region sequence-specific DNA binding transcription factor activity involved in negative regulation of transcription
GO:0000980 RNA polymerase II distal enhancer sequence-specific DNA binding	GO:0003705 sequence-specific distal enhancer binding RNA polymerase II transcription factor activity	GO:0001205 RNA polymerase II distal enhancer sequence-specific DNA binding transcription factor activity involved in positive regulation of transcription	GO:0001206 RNA polymerase II distal enhancer sequence-specific DNA binding transcription factor activity involved in negative regulation of transcription

Evidence codes for DNA binding TF activity

DNA binding (MF) and transcription regulation (BP) annotations from

- *same publication* and with *same evidence code* (either both IDA or both IMP), - DNA binding TF activity (MF) term receives this GO evidence code
- *same publication* but with *different evidence codes* (IDA and IMP), - DNA binding TF activity (MF) term is repeated twice, once with each of the two GO evidence codes
- *two different publications*: use GO evidence code 'IC: Inferred by curator'.

To generate GO evidence code 'IC':
The two GO identifiers (DNA binding and transcription regulation) assigned to the same TF from two different publications are inserted into the 'with/from' field. Reference GO_REF:0000036 is generated
(see also: http://www.geneontology.org/GO.evidence.shtml#ic)

Table 4: Evidence code table

DNA binding	Transcription regulation	TF activity
IDA	IDA	IDA/IC
IMP	IMP	IMP/IC
IDA	IMP	IDA, IMP/IC
IMP	IDA	IMP, IDA/IC

TF binding TF activity annotations

Use Table 5 to identify valid experimental evidence for one of the GO annotations for TF-binding shown in orange in Figure 6.

The IPI evidence code indicates that the interaction is a direct 1:1 interaction. The IDA evidence code should be used when the protein being annotated is shown to bind to a TF that is a complex.

The TF binding partner(s) must be recorded in '*with/from*' field (http://www.geneontology.org/GO.evidence.shtml).

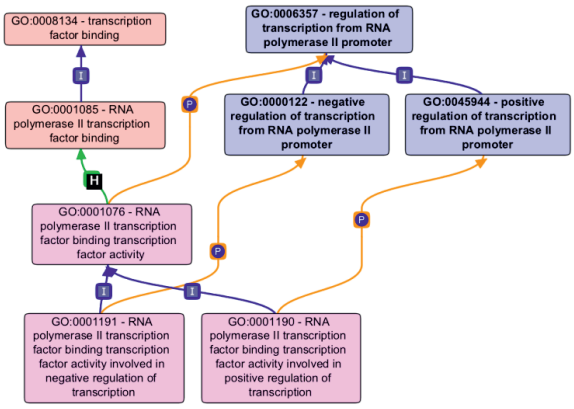


Figure 6: GO terms for MF TF binding TF activity

Transcription regulation BP annotations (shown in blue in Figure 6) are made as described above for *Transcription Regulation annotations,* page 3.

Table 5: TF binding: assays

Assays	GO evidence Code	PSI-MI code
2-hybrid interactions	IPI	MI:0018
Co-purification	IPI, IDA	MI:0004
Co-immunoprecipitation	IPI, IDA	MI:0019

Use Decision Table 6 to identify the correct GO annotation for TF-binding TF activity shown in pink in Figure 6.

For assignment of **evidence code for TF binding activity** see Table 7: when TF binding (MF) and transcription regulation (BP) annotations are from

- *same publication* but with *different evidence codes* (IDA, IPI or IMP), - TF binding TF activity (MF) term is repeated twice or three times, once with each of the GO evidence codes
- *two different publications*: use GO evidence code 'IC: Inferred by curator' (as described on page 5).

Table 6: Decision table TF binding TF activity

TF binding terms (MF)	*Transcription regulation terms (BP)*		
	GO:0006357 regulation of transcription from RNA polymerase II promoter	*GO:0045944 positive regulation of transcription from RNA polymerase II promoter*	*GO:0000122 negative regulation of transcription from RNA polymerase II promoter*
GO: 0008134 Transcription factor binding	**GO: 0001076** RNA polymerase II transcription factor binding transcription factor activity	**GO:0001190** RNA polymerase II transcription factor binding transcription factor activity involved in positive regulation of transcription	**GO:0001191** RNA polymerase II transcription factor binding transcription factor activity involved in negative regulation of transcription
GO: 0001085 RNA polymerase II transcription factor binding	**GO: 0001076** RNA polymerase II transcription factor binding transcription factor activity	**GO:0001190** RNA polymerase II transcription factor binding transcription factor activity involved in positive regulation of transcription	**GO:0001191** RNA polymerase II transcription factor binding transcription factor activity involved in negative regulation of transcription

Table 7: TF binding activity evidence code table

TF-binding	Transcription regulation	TF activity
IDA	IDA	IDA/IC
IMP	IMP	IMP/IC
IDA	IMP	IDA, IMP/IC
IMP	IDA	IMP, IDA/IC
IPI	IDA	IPI, IDA/ IC

When the functional unit of a TF is a complex

When the complex is a **homodimer, or higher order multimer of the same protein**, there are no special annotation issues as all of the activities demonstrated are properties of the same gene product.

Additional considerations for heterodimers and other multisubunit complexes

The **"contributes to" qualifier** must always be used to indicate the annotation of functions that occur in the context of complexes. The "contributes to" qualifier can be used in conjunction with any MF term, including the "transcription factor activity" terms, to indicate that it contributes to that function within the context of a complex, even though it does not possess that activity independently

In the case of a **heterodimer where one of the two proteins does not bind DNA on its own but is still found to contribute to the sequence specific binding of the other subunit** within a heterodimer: the subunit that does not bind DNA alone can still be annotated to "sequence-specific DNA binding", or possibly a more specific term, using the qualifier "contributes to" to indicate that it contributes to the DNA binding of the heterodimer.

In a **multisubunit TF where the DNA binding activity is known to be confined to one or more specific subunits**: other subunits should **not** be annotated to a "DNA binding" term.

For any subunit within a TF complex, it is appropriate to annotate to all appropriate GO terms for which that function has been experimentally shown, either individually or as part of the complex indicated with the "contributes to" qualifier. Thus, in some cases, a given protein may be annotated both a 'sequence specific DNA binding RNAP II transcription factor activity' term as well as a 'TF binding RNAP II transcription factor activity' term.

Target gene (TG) annotations

Use the 'annotation extension' column shown in Figure 7 to capture information for one or several TGs shown to be regulated by the TF whose function is annotated by using "has_regulation_target", source of TG identifier and TG identifier, as explained below.

!DB	DB Object ID	DB Object Symbol	Qualifier	GO ID	DB:Reference	Evidence Code	With (or) From	Aspect	DB Object Name	Object Synony m	DB Object Type	Taxon	Date	Assigned By	Annotation Extension
UniProtKB	O43248	HOXC11		GO:0000978	PMID:9582375	IDA		F	Homeobox protein	Hox-C11	protein	taxon:9606	20130412	NTNU_SB	
UniProtKB	O43248	HOXC11		GO:0045944	PMID:9582375	IDA		P	Homeobox protein	Hox-C11	protein	taxon:9606	20130412	NTNU_SB	has_regulation_target(UniProtKB:P09848)
UniProtKB	O43248	HOXC11		GO:0001077	PMID:9582375	IDA		F	Homeobox protein	Hox-C11	protein	taxon:9606	20130412	NTNU_SB	
UniProtKB	P47902	CDX1		GO:0000980	PMID:1577494	IDA		F	Homeobox protein	CDX-1	protein	taxon:9606	20130412	NTNU_SB	
UniProtKB	P47902	CDX1		GO:0045944	PMID:1577494	IDA		P	Homeobox protein	CDX-1	protein	taxon:9606	20130412	NTNU_SB	has_regulation_target(UniProtKB:P09923)
UniProtKB	P47902	CDX1		GO:0001205	PMID:1577494	IDA		F	Homeobox protein	CDX-1	protein	taxon:9606	20130412	NTNU_SB	

Figure 7: DbTF sample annotations with TG information (rows 2 and 5) shown as "column 16" of the spread sheet GAF2.0 format; http://www.geneontology.org/GO.format.gaf-2_0.shtml.

The relationship **''has_regulation_target''** is used to capture TG information in the spread sheet row used to record the transcription regulation term. This row can either hold terms that have is_a relationships to the term "biological regulation" (i.e. are BP-terms), or MF terms representing regulators that are part of regulatory processes (i.e. have part_of relationships to a BP regulation term).
In the examples shown in Figure 7, the transcriptiojn regulation term is GO:0045944 *positive regulation of transcription from RNA polymerase II promoter*.

When we use the "has_regulation_target" relationship, we are saying that the GO term used for the annotation, e.g. "*regulation of transcription from RNA polymerase II*" or "*sequence-specific DNA binding RNA polymerase II transcription factor activity*" has a target, and we use a gene ID (URI) to specify what that target is.

To indicate multiple TGs in the same annotation: separate each 'relationship(identifier)' pair with a pipe, "|".

Example: to capture the two TGs, the annotation extension column should contain:

has_regulation_target(source:GeneURI1) | has_regulation_target (source:GeneURI2)

where **source** can be UniProtKB, ENSEMBL, Entrez, or a model organism database, e.g. MGI, RGD, etc. and **Gene URI1** and **GeneURI2** denote identifiers from any of the above sources.

GO terms

Transcription terms

GO:0006357 regulation of transcription from RNA polymerase II promoter
Definition: “Any process that modulates the frequency, rate or extent of transcription from an RNA polymerase II promoter.”

GO:0045944 positive regulation of transcription from RNA polymerase II promoter
Definition: “Any process that activates or increases the frequency, rate or extent of transcription from an RNA polymerase II promoter.”

GO:0000122 negative regulation of transcription from RNA polymerase II promoter
Definition: “Any process that stops, prevents, or reduces the frequency, rate or extent of transcription from an RNA polymerase II promoter.”

DNA Binding terms

GO:0043565 - sequence-specific DNA binding
Definition: "Interacting selectively and non-covalently with DNA of a specific nucleotide composition, e.g. GC-rich DNA binding, or with a specific sequence motif or type of DNA e.g. promotor binding or rDNA binding."

GO:0000977 - RNA polymerase II regulatory region sequence-specific DNA binding
Definition: "Interacting selectively and non-covalently with a specific sequence of DNA that is part of a regulatory region that controls the transcription of a gene or cistron by RNA polymerase II."

GO:0000978 - RNA polymerase II core promoter proximal region sequence-specific DNA binding
Definition: "Interacting selectively and non-covalently with a sequence of DNA that is in cis with and relatively close to a core promoter for RNA polymerase II." comment: Note that the phrase "upstream activating sequence", or UAS is often used in S. cerevisiae literature to refer to regulatory sequences that occur in the region upstream and proximal to the core promoter. In contrast, in bacteria such as E. coli, the phrase "upstream activating sequence", or UAS is a synonym for "enhancer".

GO:0000980 - RNA polymerase II distal enhancer sequence-specific DNA binding
Definition: "Interacting selectively and non-covalently with a RNA polymerase II (Pol II) distal enhancer. In mammalian cells, enhancers are distal sequences that increase the utilization of some promoters, and can function in either orientation and in any location (upstream or downstream) relative to the core promoter."

Sequence-specific DNA binding transcription factor activity terms

GO: 0000981 - sequence-specific DNA binding RNA polymerase II transcription factor activity
Definition: Interacting selectively and non-covalently with a specific DNA sequence in order to modulate transcription by RNA polymerase II. The transcription factor may or may not also interact selectively with a protein or macromolecular complex.
GO:0001227 - RNA polymerase II transcription regulatory region sequence-specific DNA binding transcription factor activity involved in negative regulation of transcription
Definition: Interacting selectively and non-covalently with a sequence of DNA that is in the regulatory region for RNA polymerase II (RNAP II) in order to stop, prevent, or reduce the frequency, rate or extent of transcription from an RNA polymerase II promoter.

GO:0001228 - RNA polymerase II transcription regulatory region sequence-specific DNA binding transcription factor activity involved in positive regulation of transcription
Definition: Interacting selectively and non-covalently with a sequence of DNA that is in the transcription regulatory region for RNA polymerase II (RNAP II) in order to activate or increase the frequency, rate or extent of transcription from the RNAP II promoter.

GO:0000982 - RNA polymerase II core promoter proximal region sequence-specific DNA binding transcription factor activity
Definition: Interacting selectively and non-covalently with a sequence of DNA that is in cis with and relatively close to a core promoter for RNA polymerase II (RNAP II) in order to modulate transcription by RNAP II.

GO:0001077- RNA polymerase II core promoter proximal region sequence-specific DNA binding transcription factor activity involved in positive regulation of transcription
Definition : Interacting selectively and non-covalently with a sequence of DNA that is in cis with and relatively close to a core promoter for RNA polymerase II (RNAP II) in order to activate or increase the frequency, rate or extent of transcription from the RNAP II promoter.

GO:0001078- RNA polymerase II core promoter proximal region sequence-specific DNA binding transcription factor activity involved in negative regulation of transcription
Definition: Interacting selectively and non-covalently with a sequence of DNA that is in cis with and relatively close to a core promoter for RNA polymerase II (RNAP II) in order to stop, prevent, or reduce the frequency, rate or extent of transcription from the RNAP II promoter.

GO:0003705 - sequence-specific distal enhancer binding RNA polymerase II transcription factor activity
Definition: Interacting selectively and non-covalently with a sequence of DNA that is in a distal enhancer region for RNA polymerase II (RNAP II) in order to modulate transcription by RNAP II.

GO:0001205 - **RNA polymerase II distal enhancer sequence-specific DNA binding transcription factor activity involved in positive regulation of transcription**.
Definition: "Interacting selectively and non-covalently with a sequence of DNA that is in a distal enhancer region for RNA polymerase II (RNAP II) in order to activate or increase the frequency, rate or extent of transcription from the RNAP II promoter."

GO:0001206 - **RNA polymerase II distal enhancer sequence-specific DNA binding transcription factor activity involved in negative regulation of transcription**.
Definition: "Interacting selectively and non-covalently with a sequence of DNA that is in a distal enhancer region for RNA polymerase II (RNAP II) in order to stop, prevent, or reduce the frequency, rate or extent of transcription from an RNA polymerase II promoter."

Transcription factor binding terms

GO:0008134 - Transcription factor binding (MF)
Definition: Interacting selectively and non-covalently with a transcription factor, any protein required to initiate or regulate transcription.

GO:0001085 - RNA polymerase II transcription factor binding (MF)
Definition: Interacting selectively and non-covalently with an RNA polymerase II transcription factor, any protein required to initiate or regulate transcription by RNA polymerase II.

Transcription factor binding transcription factor activity terms

GO:0001076 - RNA polymerase II transcription factor binding transcription factor activity (BP)
Definition: Interacting selectively and non-covalently with an RNA polymerase II transcription factor, which may be a single protein or a complex, in order to modulate transcription. A protein binding transcription factor may or may not also interact with the template nucleic acid (either DNA or RNA) as well.

GO:0001190 - RNA polymerase II transcription factor binding transcription factor activity involved in positive regulation of transcription (BP)
Definition: Interacting selectively and non-covalently with an RNA polymerase II transcription factor, which may be a single protein or a complex, in order to increase the frequency, rate or extent of transcription from an RNA polymerase II promoter. A protein binding transcription factor may or may not also interact with the template nucleic acid (either DNA or RNA) as well.

GO:0001191 - RNA polymerase II transcription factor binding transcription factor activity involved in negative regulation of transcription (BP)
Definition: Interacting selectively and non-covalently with an RNA polymerase II transcription factor, which may be a single protein or a complex, in order to stop, prevent, or reduce the frequency, rate or extent of transcription from an RNA polymerase II promoter. A protein binding transcription factor may or may not also interact with the template nucleic acid (either DNA or RNA) as well.

GOC Evidence Codes

IDA – Inferred from direct assay
Description (GOC): The IDA evidence code is used to indicate that a direct assay was carried out to determine the function, process, or component indicated by the GO term.

IMP – Inferred from mutant phenotype
Description (GOC): The IMP evidence code covers those cases when the function, process or cellular localization of a gene product is inferred based on differences in the function, process, or cellular localization between two different alleles of the corresponding gene. The IMP code is used for cases where one allele may be designated 'wild-type' and another as 'mutant'. It is also used in cases where allelic variation occurs naturally and no specific allele is designated as wild-type or mutant.

IC - Inferred by Curator
Description (GOC): The IC evidence code is to be used for those cases where an annotation is not supported by any direct evidence, but can be reasonably inferred by a curator from other GO annotations, for which evidence is available.

IPI - Inferred from Physical Interaction
Description (GOC): Covers physical interactions between the gene product of interest and another molecule (such as a protein, ion or complex). IPI can be thought of as a type of IDA, where the actual binding partner or target can be specified, using "with" in the with/from field.

Paper IV

Is not included due to copyright