CONDITION-SPECIFIC DIFFERENTIAL SUBNETWORK ANALYSIS FOR BIOLOGICAL SYSTEMS

Deepali Jhamb

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the School of Informatics and Computing,
Indiana University

March 2015

Accepted by the Graduate Faculty, Indiana University, in partial
fulfillment of the requirements for the degree of Doctor of Philosophy.

Xiaowen Liu, PhD, Chair

Mathew J. Palakal, PhD

Doctoral Committee

David L. Stocum, PhD

December 12, 2014

Lang Li, PhD

Yunlong Liu, PhD

DEDICATION

I dedicate my dissertation work to my parents who taught me to believe in myself and supported me to follow my passion for science. I also dedicate this work to my brother who made me laugh even in the toughest situations. I especially thank and dedicate this work to my dear husband for his endless love, support, and encouragement throughout this journey.

ACKNOWLEDGEMENTS

The last few years to get my PhD have been quite a journey. I learned many new skills, both professional and personal, and strengthened my interest in science and research. Most of all, this degree taught me to be patient and not give up!

There are many people who have supported me through this work and I would like to thank them all. Primarily, I would like to thank my *gurus* Dr. Mathew J. Palakal and Dr. David L. Stocum for their guidance and continual support. Not only did they encourage me to develop innovative methods for biological data analysis, they also taught me great interpersonal skills. Their enthusiasm and passion for research has been truly inspirational. I would also like to thank all my other committee members, Dr. Lang Li, Dr. Yunlong Liu, and Dr. Xiaowen Liu for their valuable time and suggestions that have helped improve this work.

I especially thank all the members of the TiMAP lab, particularly Dr. Meeta Pradhan. Meeta has been a great friend in the lab and her in-depth review of the work and manuscripts has been invaluable to my research. Stocum lab members have also contributed greatly to my work: a big thanks to Nandini Rao, Fengyu Song, and Xiaoping Chen for helping me validate the Informatics hypothesis and in writing the manuscripts. I would also like to acknowledge our collaborators at the University of Illinois, Urbana-Champaign, Dr. Jo Ann Cameron, and Derek Milner for their help with validation of the proteomics datasets.

Many others in the IUPUI community have been instrumental during this work. I thank the amazing faculty and staff at the School of Informatics and Computing. Mary O'Neill, Jeff Hostetler, and Elizabeth Bunge helped me through all the administrative work. A special thanks to Jeff for proofreading this dissertation, it was a big help! I also thank Debra Barker at the Graduate School who made sure that all my documents were submitted in time and reviewed the format of this dissertation.

It is difficult to express my gratitude for my family and the wonderful bunch of friends who have stood by me at all times. I have been blessed to have you all in my life. Heartfelt thanks to Rajendra Kedlaya, Sreejith Ramakrishnan, Abhinita Mohanty, and Sudha Savant. Finally, I thank the Almighty for giving me such an amazing opportunity.

This work was supported by grants from the W.M. Keck Foundation, the US Army Research Office (#W911NF07-10176), and the Telemedicine & Advanced Technology Research Center (TATRC, W81XWH-11-1-0743) to David L. Stocum.

My research is an important part of my life and I hope it will make a big impact in the lives of others someday.

Deepali Jhamb

CONDITION-SPECIFIC DIFFERENTIAL SUBNETWORK ANALYSIS FOR BIOLOGICAL SYSTEMS

Biological systems behave differently under different conditions. Advances in sequencing technology over the last decade have led to the generation of enormous amounts of condition-specific data. However, these measurements often fail to identify low abundance genes/proteins that can be biologically crucial. In this work, a novel text-mining system was first developed to extract condition-specific proteins from the biomedical literature. The literature-derived data was then combined with proteomics data to construct condition-specific protein interaction networks. Further, an innovative condition-specific differential analysis approach was designed to identify key differences, in the form of subnetworks, between any two given biological systems.

The framework developed here was implemented to understand the differences between limb regeneration-competent *Ambystoma mexicanum* and –deficient *Xenopus laevis*. This study provides an exhaustive systems level analysis to compare regeneration competent and deficient subnetworks to show how molecular entities inter-connect with each other and are rewired during the formation of an accumulation blastema in regenerating axolotl limbs. This study also demonstrates the importance of literature-derived knowledge, specific to limb regeneration, to augment the systems biology analysis. Our findings show that although the proteins might be common between the two given biological conditions, they can have a high dissimilarity based on their biological and topological properties in the subnetwork. The knowledge gained from the distinguishing features of limb regeneration in amphibians can be used in future to chemically induce regeneration in mammalian systems.

The approach developed in this dissertation is scalable and adaptable to understand differential subnetworks between any two biological systems. This methodology will not only facilitate the understanding of biological processes and molecular functions which govern a given system but also provide novel intuitions about the pathophysiology of diseases/conditions.

Xiaowen Liu, PhD, Chair

TABLE OF CONTENTS

List of Tables

List of Figures

List of Abbreviations

AEC	Apical Epithelial Cap
ANOVA	Analysis Of Variance
AP	Anteroposterior
ATM	Automatic Term Mapping
AUC	Area Under the Curve
BLASTp	Basic Local Alignment Search Tool for Proteins
BP	Biological Process
CL	Concept List
CSD	Critical Size Defect
CSDB	Condition Specific Database
CV	Coefficient of Variance
Dpa	Days Post Amputation
DS	Dissimilarity Score
EBI	European Bioinformatics Institute
ECM	Extracellular Matrix
EFO	Experimental Factor Ontology
EHDA	Human Developmental Anatomy Ontology
Eq	Equation
EST	Expressed Sequence Tags
FC	Fold Change
FDR	False Discovery Rate
FN	False Negative
FP	False Positive
FPR	False Positive Rate
GF	Growth Factor
GO	Gene Ontology
HPLC	High Performance Liquid Chromatography
HPRD	Human Protein Reference Database
iHOP	Information Hyperlinked Over Proteins

IPA	Ingenuity Pathway Analysis
IR	Information Retrieval
KEGG	Kyoto Encyclopedia of Genes And Genomes
KPGP	Korean Personal Genome Project
LC	Liquid Chromatography
MCC	Mathew’s Correlation Coefficient
McSyBi	Multi-Document Clustering System for Biomedicine
MeSH/MESH	Medical Subject Headings
MF	Molecular Function
MS	Mass Spectrometry
NCBI	National Center for Biotechnology Information
NCBO	National Center for Biomedical Ontology
NCIT	National Cancer Institute Thesaurus
NF	Nieuwkoop-Faber
NGS	Next Generation Sequencing
NIFSTD	Neuroscience Information Framework Standard Ontology
No.	Number
OMIM	Online Mendelian Inheritance In Man
PBS	Phosphate Buffered Saline
PIDM	Protein Interaction detection Method
PMID	Pubmed Identifier
PMRA	Pubmed Related Article
PPIs	Protein Protein Interactions
PubNet	Publication Network Graph Utility
RankSVM	Rank Support Vector Machines
RefMed	Relevance Feedback Search Engine for Pubmed
RegEx	Regular Expressions
ROC	Receiver Operating Characteristic
SIS	Small Intestine Submucosa
SNOMED CT	Systematized Nomenclature Of Medicine - Clinical Terms
SQL	Structured Query Language

tBLASTn	Translated BLAST
TF	Transcription Factor
TF-IDF	Term Frequency-Inverse Document Frequency
TN	True Negative
TP	True Positive
TPR	True Positive Rate
TREC	Text Retrieval Conferences
UMLS	Unified Medical Language System
XLS	Excel
XML	Extensible Markup Language

CHAPTER ONE: INTRODUCTION

Overview of the Problem

Systems biology was introduced in 2001 as a framework to study the behavior and relationships between different entities of a biological system [1]. The last fourteen years have seen tremendous progress in this field, leading to a paradigm shift in biology—from being a descriptive science to a predictive science. In the 20th century, reductionism dominated the research in biology. It was based on the "divide and conquer" policy and hence focused on identification of smaller, simpler solutions of a complex biological system. This approach was largely successful in providing information about several biological processes and molecular functions and also resulted in finding cures for several diseases, especially metabolic disorders. However, for complex diseases such as cancer, this approach has not succeeded and as more data is being collected, it has become clear that the complexity of a biological system is greater than the mere sum of its individual parts [1-4].

With the advent of sequencing technologies, enormous amounts of data are being generated and deposited in public repositories such as the Gene Expression Omnibus [5]. High throughput measurements are an important source of large and heterogeneous biological information such as genomics, transcriptomics, proteomics, interactomics, and variant data. However, high throughput measurements often fail to identify low abundant genes/proteins which can be biologically crucial [6, 7]. Biomedical literature or publications are the most comprehensive resource of the knowledge amassed in this discipline. This collection can hence be used to extract the "missing" knowledge (not obtained by high throughput analysis). Since it is hard for scientists to manually keep up with this exploding amount of biomedical literature, text mining or automated retrieval of knowledge from biomedical literature has gained more importance over the last few years. Text mining has its strengths and weaknesses, the most common weakness being the noisy and unspecific data generated as a result of natural language processing. It is important to develop effective and efficient information retrieval techniques to circumvent the problem of noisy data.

An important issue that needs to be addressed during automated information retrieval is content-focused querying. Any particular biological entity is often discussed in sufficient detail in a given article. However, the traditional text mining approaches usually discard these content-specific details and return a subset of documents which contain the keywords specified by the user. This results in generation of a lot of false-positives or noisy data. For example, from a sentence like “collagen deposition is suppressed during limb regeneration, so we investigated collagen deposition and apical epithelial cap (AEC) formation during axolotl limb regeneration,” biological features such as appendage (limb), biological process (regeneration), tissue (AEC), and organism (axolotl) are usually ignored by keyword retrieval methods (unless these were specified as keywords). Such biological features are crucial to defining and understanding the concept/context of an article. Several document clustering or text categorization approaches have been used to cluster the documents based on context. However, these approaches either suffer from the problem of high dimensionality and hence cannot be applied to large collections such as PubMed [8, 9] or are based on ontologies such as Medical Subject Headings (MeSH) [10] which are not always up to date and can generate a high number of false positives. The performance evaluation of these methods has suggested that it varies based on the domain/field under investigation [8].

Systems biology models can be used to logically integrate the knowledge (such as genes/proteins) extracted from context-specific text mining and high throughput measurements. This approach can build the foundation to integrate several datasets in an attempt to better understand complex biological systems. Network analysis has found application in several areas ranging from electrical circuits to social networks. It is also now being extensively used to study the inter-relations between different molecular components of a biological system [11]. Most of the work in network biology has focused on static networks. Static networks however merely represent the state of a system at any given point and cannot be used to make predictions about network behavior. As described by Hiroki Kitano [4], one of the pioneers in systems biology:

“*Although such a diagram represents an important first step, it is analogous to a static roadmap, whereas what we really seek to know are the traffic patterns, why such traffic patterns emerge, and how we can control them.*”

One of the key challenges then is to understand how the networks change with different states of a system. In other words, it is critical to understand the network dynamics or “rewiring.” Biological systems behave differently under different conditions and network comparisons such as between disease and normal state can offer novel intuitions into the pathophysiological process of a disease and also suggest biomarkers/drug targets for the same. This can also help formulate a novel hypothesis about the change in the biological processes, and regulation patterns across different conditions.

Limitations of the Present Approaches

Several studies have been performed for context-specific mining and network comparisons in the biomedical domain (as described in the Background section of this thesis). However, the following are some of the pitfalls in current studies:

1. Most importantly, to our knowledge none of the studies so far has logically integrated context-specific text mining and high throughput datasets in a systems biology framework to compare subnetworks across different biological conditions for target discovery. Most of the studies have either focused on information retrieval techniques or systems biology analysis on high throughput measurements alone. In our opinion, information retrieval, information extraction, and downstream analysis through systems biology together can help formulate biologically meaningful hypotheses to identify new targets or to understand the pathophysiology of complex biological systems.
2. The present work in document clustering or context based information retrieval suffers from the “curse of dimensionality” and hence cannot be used to cluster a large set of documents. The current approaches also generate a lot of false positives and the performance of a given method is highly dependent on the field/domain.
3. Although a great deal of work has been done in the identification of functional modules and network analysis, not enough work has been done to identify functionally differential components by comparing subnetworks/networks between biological systems. The current network comparison algorithms are focused on either the local or global alignment of networks with respect to sequence- or structure-based similarity between the network nodes. These approaches are suitable to find the common structures between networks so as to establish phylogenetic relationships but

are not capable of identifying the functionally differential subnetwork components between the two conditions.

4. Network directionality has not been considered in most of the current network comparison approaches, which is critical to understanding the regulatory mechanisms and both the downstream and upstream effectors of a process.
5. Existing tools like Ingenuity Pathway Analysis (IPA) [12], MetaCore^{TM} from GeneGO [13] also perform enrichment analysis, however their enrichment is restricted to the use of either the function or pathway information and they do not integrate condition-specific data from the biomedical literature. These tools have built-in algorithms to generate a list of important networks in a biological condition but do not contain effective differential network comparison algorithms. Moreover, these are commercial software tools and not freely available public tools.
6. Network comparisons have been done based on either topology or coexpression networks from high throughput datasets. Each has its own limitations—for instance, topology alone might not be able to identify biologically relevant information while high throughput datasets often fail to identify genes/proteins with relatively low expression level.

Summary of the Methodology

The methodology for the systematic subnetwork comparison between biological conditions developed in this study can be broadly categorized into three steps: integration, filtration, and analysis. Briefly, an innovative algorithm was designed to mine the condition-specific (or context-specific) knowledge from the biomedical literature which and integrate it with the high throughput data. This information was used to construct protein interaction networks which were filtered using a rule based novel algorithm to generate subnetworks, based on both the topological (interaction profile) and biological parameters such as molecular class, expression, literature relevance, function and pathway information. A unique subnetwork comparison algorithm to identify differential subnetworks then analyzed subnetworks between two conditions. This algorithm, unlike current approaches, also considered the direction of interaction between the proteins on the subnetworks.

This framework described above was implemented to understand the difference between the limb regeneration competent system of *Ambystoma mexicanum* (axolotl) and the limb regeneration deficient system of adult *Xenopus laevis* froglets. Proteomics data from the amputated limbs of both systems at multiple time points was used as the high throughput data in this study. The context-specific information retrieval method was then used to retrieve regeneration-specific articles. Proteins extracted from these relevant articles and proteomics data proteins were used to build protein interaction networks for the limb regeneration-competent and -deficient systems. The novel subnetwork comparison algorithm was further used to identify the most differential growth factor (GF), transcription factor (TF) and extracellular matrix (ECM) protein subnetworks in these conditions. This led to the generation of a hypothesis to suggest potential protein targets which can be instrumental in conferring limb-regeneration ability on *Ambystoma mexicanum*. A similar condition-specific data mining methodology was applied in order to understand the segment defect regeneration across a critical size defect. The key GFs identified by this study were validated in the biology laboratory and yielded positive results, thus demonstrating the efficiency of the approach developed.

The approach developed in this dissertation is scalable and adaptable to understand differential subnetworks between any two biological systems. This methodology will not only facilitate the understanding of biological processes and molecular functions which govern a given system but also provide novel intuitions about the pathophysiology of diseases/conditions.

Significance

This approach is expected to increase general understanding about the underlying mechanisms of a biological condition. As an example, a comparison between the regeneration-competent and –deficient system provided insight into the complex mechanisms which confer regeneration ability on urodeles as compared to adult anurans. The knowledge gained from the distinguishing features of limb regeneration at systems level in amphibians can then be used to chemically induce regeneration in mammalian systems. These mechanisms can also be further analyzed to understand why humans are not capable of regenerating complex tissues. Hence, significant targets might be identified which in future can be used to confer this ability in humans. Similar analysis can also be performed to distinguish other biological systems—such as performing a comparison between different types of cancer in order to provide intuitions about the pathophysiological processes of cancer.

Innovation

1. The novel bibliomics methodology proposed in this study identified context-specific data from the biomedical literature. Traditional text mining approaches lack in the identification of large-scale context-specific information from the text. This approach was used to identify limb regeneration-specific articles from a set of approximately 200,000 documents—which is higher than any of the currently available methods. The approach developed here achieved a Mathew's Correlation Coefficient (MCC) score of 0.92 and is easily scalable to a much higher number of articles.
2. An innovative rule based algorithm was developed to identify subnetworks from a global context-specific biological network. This algorithm used both the biological and topological properties of proteins on the network.
3. To our knowledge, this is the only functional and molecular class based differential subnetwork analysis which includes directed protein interactions between proteins mined from condition-specific literature and high throughput experiments.
4. Systems biology approaches have not been applied yet to study the limb regeneration system and the present studies have not been able to confer regeneration-competence upon the deficient mammals. We propose a novel way to analyze the regeneration system so as to discover the governing molecules and mechanisms of limb regeneration in axolotl.
5. In a screen of growth factor combinations (identified by text mining and systems biology) and protein extracts of axolotl whole limb and regeneration blastema tissues, we found that a combination of BMP-4 and HGF, as well as limb tissue protein extract, but not blastema extract of amputated limbs, stimulated skeletal regeneration across 50% segment defects when delivered by a pig small intestine submucosa (SIS) scaffold.
6. There is enough evidence to indicate that regenerating cells show stem cell-like characteristics. Hence, we believe that the regeneration mechanisms unraveled here will also help in the progression of stem cell research and medicine.

CHAPTER TWO: BACKGROUND

Bibliomics

Text mining refers to the discovery of novel patterns by automatically extracting information from text. This usually involves “connecting the dots” or different pieces of information from text and linking them together to derive a meaningful, testable hypothesis [14]. Within the domain of text mining, the bioinformatics discipline which deals specifically with the structural and semantic analysis of the vast biomedical literature is generally referred to as *bibliomics*. The advancement of new high throughput technologies and research capabilities in the last decade have contributed to the exponential growth of biomedical literature [9]. As a result, bibliomics has gained more importance over the last few years. The ultimate goal of bibliomics is to make relevant judgments about new targets to help in progression of basic science and drug discovery.

Biomedical Literature

Biology has often been referred to as a knowledge-based science—unlike chemistry and physics that can be defined by laws and mathematical equations. There is a huge amount of biological data (especially with the advent of current genomics platforms) available for the research community. However, this data is complex, volatile, and heterogeneous [15, 16]. PubMed, a resource developed and maintained by the National Center for Biotechnology Information (NCBI), provides free access to over 24 million citations and abstracts in the biomedical literature [17]. In addition to PubMed there are several other data resources available from NCBI and the European Bioinformatics Institute (EBI) for biomedical and genomic information [18, 19]. Several other initiatives have been taken by different research groups and institutions for organization of biological information such as, but not limited to, UniProt [20], the Human Protein Reference Database (HPRD) [21], and BioGRID [22].

Information Retrieval

Information retrieval (IR) or obtaining relevant information from a large collection is typically the first task in text mining. Although PubMed provides the interface for querying the most comprehensive resource for the biomedical literature, users are often overwhelmed with the long list of results. It has been previously shown that over one-third of the searches on PubMed yield 100 or more documents [23]. The strategy for IR in PubMed uses automatic term mapping (ATM) for a given query. Briefly, the terms in a query are searched in several lists in the following order: MeSH terms, journal names, and author names. If there is a match in any list, mapping stops and the matched terms and query terms are searched in the "All Fields" of PubMed. If there is no match, PubMed builds a Boolean query with the query terms and searches in the "All Fields" [24]. Query expansion methods like the one used in PubMed and other methods such as retrieval feedback aim to increase the search accuracy [25]. However, some research suggests an improved performance while others suggest no improvement in retrieval by using the query expansion methods [26]. The accuracy of these methods varies with topics, the most common issue being the increased number of false positives. Since the users searching PubMed are usually interested in more specific hits to their domain, the value of query expansion to the end user is questionable [27].

Document Similarity

Document similarity is very crucial for the purpose of IR and several methods have been investigated which aim to improve the ranking of the search results so that similar documents (in relation to the query) rank higher [28, 29]. Text REtrieval Conferences (TREC) have been a significant part of this effort including their genomics track which ran from 2003-2007 [30, 31]. The methods used for ranking the retrieved documents are generally derived from two earlier publications on vector space models or probabilistic models [28, 32]. MedlineRanker uses supervised learning to identify the list of discriminative words (nouns in a set of known documents) by using a linear Naïve Bayesian classifier. This method can rank a maximum of the 10,000 most recent articles in PubMed, related to the initial set of known domain-specific documents [33]. Several other features such as MeSH terms, Unified Medical Language System (UMLS) concepts, gene ontology (GO) terms, author names, journals, year of publication have been used by tools such as XplorMed, Multi-document clustering system for Biomedicine (McSyBi), and GOPubmed to improve the performance of IR in biomedical literature [34-37]. Alibaba, PubMed-EX, Information Hyperlinked over Proteins (iHOP), and EBIMed identify biomedical entities such as proteins, genes, drugs, diseases, etc. and then compute their co-occurrence in a sentence [38-41]. The method employed in Publication Network Graph Utility (PubNet) uses articles, authors, genes, or MeSH terms or location as nodes on the network [42]. The Relevance Feedback Search Engine for PubMed (RefMed) and MiSearch are examples of methods based on user feedback [43, 44]. RefMed uses learning to rank algorithms and Rank Support Vector Machines (RankSVM) to learn relevant documents based on user feedback [43]. The different methods to enhance IR in biomedical literature have been summarized effectively by Zhiyong Lu [9].

The different document similarity approaches in the biomedical domain can be broadly classified into: text-based, citation-based, and hybrid approaches [8]. The text-based approaches use a natural language processing toolkit such as tf-idf (term frequency-inverse document frequency) to identify parts of text (words or phrases) which can be used for document clustering. Citation-based approaches are based on the concept that similar articles must have similar bibliographic information. Hybrid approaches integrate both text and citation -based methods. Performance results of these approaches are domain/field specific and hence there is no consensus on the best performing method.

Document Clustering

Document clustering or text categorization groups similar documents to provide relevant results for a users' query. Document clustering is often limited to a small set of documents since it suffers from the "curse of dimensionality." Most of the approaches for document clustering follow a vector space model to represent the important words from the text and hence suffer from the problem of high dimensionality [45]. Several different machine learning or statistical approaches have been employed for document clustering such as Bayes probability distribution [46], neural networks [47], nearest neighbor classification [48], decision trees [49], etc.

One of the first and most well recognized studies that described document clustering on biomedical data, TextQuest, was based on creating a fixed length array bit vector representation of important words (identified by tf-idf and frequency distribution) for each document and then using k-means unsupervised clustering to cluster the documents. A final set of representational terms of a cluster were identified by using log-odds ratio [50]. This work has been extended as BioTextQuest and more recently as BioTextQuest^{+} [51, 52]. Although the fundamental principles are the same as in the initial study, it uses better stemming and clustering algorithms. In the most recent version, queries can be made on both PubMed and Online Mendelian Inheritance in Man (OMIM) resources—it also enables extraction of biological entities such as genes and proteins from the relevant document clusters. The recent versions also support better visualization of results such as tag clouds of overrepresented terms [53]. Several other approaches have taken advantage of ontologies such as MeSH to cluster articles in the biomedical literature [54-57]. The use of ontologies reduces the burden of dimensionality as well as proves to be an effective way to deal with synonyms. However, most of these approaches work only on a small set of documents and results vary based on the field/domain being investigated. The study by Boyack et al is the most comprehensive evaluation of different text-based document similarity measures for document clustering such as tf-idf, latent semantic analysis, topic modeling, self-organizing maps, and Poisson-based methods [8]. In their study, they found PubMed Related Article search algorithm (PMRA) [58] and BM25 [59] perform the best.

The PubMed Related Article (PMRA) search algorithm and BM25, both use Poisson distribution for calculating term-weight. However, there are fundamental differences between these two approaches. The main goal of PMRA is to identify "relatedness" of documents rather than "relevance" as estimated by BM25. The retrieval model employed in PMRA is used to populate the related articles in the right hand panel of PubMed. The derivation of Poisson parameters in these two models has important differences especially in the definition of "elite" vs. "non-elite" distributions. PMRA method also accounts for document length since identification of related documents uses the entire document as a query and so query length normalization plays a significant role in the model. PMRA is a topic based content similarity model which uses MeSH terms, title, and abstract words to determine concept based term frequencies [58, 60].

Biomedical Ontologies

Ontologies provide a computational framework to help in the structural and semantic classification of biological data. Ontologies help tremendously in the storage, representation, and dissemination of biological information [15, 16]. Some well-known ontologies in the biomedical domain are: GO[61], UMLS [62], MeSH [10], and Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT) [63]. The National Center for Biomedical Ontology (NCBO) maintains a repository of biomedical ontologies that can be accessed programmatically or through a web-based resource called BioPortal [64, 65]. A similar effort is the ZOOMA, which uses the Experimental Factor Ontology (EFO) available from EBI [66].

NCBO BioPortal provides the most comprehensive collection of biomedical ontologies. At present, it contains 393 ontologies with 5,299,586 classes. This can be easily used to build tools and web services to solve specific problems in the biomedical domain. Such ontologies facilitate translational work in biomedicine by allowing a semantic integration of biological terms. Ontologies make it possible to draw correlations between different biological entities such as diseases, proteins, etc.

Ontologies in BioPortal can be searched in many different ways, including the standard tree based search, or queried programmatically through web services [64]. The "Mapping" functionality allows the user to search if there are shared terms between two or more ontologies to facilitate direct comparisons. The "Recommender" and "Annotator" are two very useful tools for natural language processing of the biomedical text. The "Recommender" service takes the text documents (such as abstracts of articles) or keywords as input and suggests the most appropriate ontologies related to the text. It also provides a ranking of the ontologies based on coverage, connectivity, and the number of concepts in ontologies [67]. NCBO "Annotator" is the most widely used tool which takes the text as input and returns matched terms from ontologies. The user can select direct or hierarchical matching to the ontology concepts and the ontologies to use can be specified as well [68]. NCBO "Resource Index" is another useful tool which relates the ontology concepts with the metadata from the online data resources such as ArrayExpress. It can be linked with the output of Annotator to identify useful online resources related to the matching ontology concepts from the text [69].

The condition-specific data mining algorithm developed in this study used the ontologies specific to a biological domain to first build a list of concept terms. These terms were then used to retrieve condition-specific documents from which proteins were extracted and classified as literature-derived condition specific proteins. We believe that the work done here is scalable to multiple biological domains, is more specific (concept list is different and consists of terms specific for each biological domains since it is built on a known set of positive articles from that domain), and is capable of processing a much higher number of documents as compared to the current studies.

Systems Biology

Topology of Biological Networks

A biological network is modeled as a directed or undirected graph, with a set of nodes (proteins, genes, etc.) and edges (interactions between the nodes). These graphs can be weighted or unweighted on both nodes and edges. Different network analysis algorithms are now available to understand biological networks. Most of these analysis algorithms are based on topological parameters such as degree, eccentricity, closeness, radiality, etc. Degree is the number of nodes directly connected to a given node and so a node with high degree has many connections and many researchers use it as a measure by which to assign significance. Eccentricity refers to the reciprocal of the direct path from a given node to the farthest away node on the network and so a high eccentricity value for a node implies that it is close to all other nodes. Closeness, like eccentricity defines how close a node on the network is to all the other nodes. Radiality compares the direct paths of a node with the longest direct path in the network which is the network diameter. A high value of radiality implies how central a node is in the network [4, 70-76]. It has been established that biological networks are scale-free networks and not random. In scale-free networks, the degree of a network approximates to a power law $(P(k) \sim k^{-\gamma})$ where γ is found to be less than three. In other words, biological networks have a small number of hub nodes and most of the other nodes have fewer connections[77].

Topological properties can be analyzed at both local and global levels. The local topological measures only consider the direct neighborhood of a node while global measures consider the entire network. Cytoscape is free network visualization software which provides access to several plugins capable of estimating topological scores for the network. The latest version of Cytoscape (3.1) has a built-in feature, "Analyze Networks," which can be used to perform simple topological calculations [78]. CytoHUBBA is an example of a well-known Cytoscape plugin which has the capability to analyze both local and global measures of topology [79].

Topological parameters usually vary in their capability to identify essential genes. Even though the identification of essential genes worked in organisms like yeast, they are much less complex. The key molecules (or essential nodes) identified by such parameters usually tend to be hub nodes (high degree nodes) which are not necessarily biologically significant. Hub nodes might not be specific to a condition since most of them are found to be the regulators of multiple processes even in a normal state. Moreover, targeting such genes also increases the lethality of an organism and hence they cannot be used as potential key targets to induce a response or even as drug targets. It is important to infuse biological knowledge into node ranking so that the condition-specific targets can be identified. Recently, Dezso et al focused on incorporating the connectivity profile with biologically relevant information (such as gene expression) to identify important nodes in the network. Their approach ranked the nodes based on the connectivity of a given node with genes from expression data. The hub nodes that were not connected to the genes from expression were given a low score despite the large number of connections. The nodes which connect most of the genes from expression data were ranked higher by this methodology [80].

Modules and Motifs

Biological networks can be analyzed in a top down fashion or a bottom up fashion. The latter involves the use of subgraphs for the analysis and interpretation of biological data. It is known that several biological networks have a high clustering coefficient which in turn indicates presence of motifs [76, 81]. Motifs are statistically overrepresented, highly interconnected subgraphs with a distinct pattern such as triangular motifs, which form the feed forward loops in the regulatory networks [82-85]. It has also been observed that these motifs are evolutionarily conserved such as in the yeast protein interaction network [86-88]. Many other motifs have been studied such as autoregulation, single input module, dense overlapping regulons, and feedback loops [89, 90].

A module is a type of subgraph (or subnetwork) with a highly interconnected group of nodes which work together to result in a definite function [76]. Modules in biology were first studied using the genomics or proteomics data where genes/proteins were clustered based on their expression and hence coexpression modules were generated [91-93]. It became clear that these coexpression modules were also functionally related and several studies have been done to identify functional modules in the data based on several properties such as network topology, phenotypic data, expression, gene ontology, pathway information, etc. [94-100]. Modules are known to be present in metabolic, protein-protein interaction, signaling and regulatory networks. However, it should be noted that module detection methods are highly ambiguous and predict different modules for the same dataset [76]. Often times these module detection algorithms also miss properties of a network, such as directionality.

Network Comparison

Network alignment aims to find a common subgraph among the input networks. Like sequence alignment, network alignments can be used to establish evolutionary relationships, ortholog predictions, annotating the protein functions for relatively less-studied species, and to understand the biological processes in a cell. The network alignment problem can be either defined as a local or global network alignment. Local network alignment is commonly used to find regions or small subnetworks or pathways that are conserved in species. It can also refer to identification of modules with high functional similarity across species. Global network alignment on the other hand aims to find the region of maximal similarity among the given networks. Global alignment aims to map every node in the smaller network uniquely with exactly one node in the larger network. Although the local alignments create ambiguity since one node can map to more than one node in other networks, local alignments are more consistent in identifying functionally conserved regions of similarity between species and are computationally more tractable [101, 102].

Coexpression networks coupled with several statistical techniques such as singular value decomposition, Pearson correlation or biological datasets such as phenotypic, transcription regulation, and promoter data have also been extensively used for the purpose of network comparison [103-110]. Coexpression networks are believed to identify groups of differentially regulated genes or modules usually belonging to a particular biological process or function [111-113]. One of the most fundamental works in this domain by Segal et al was to identify coregulated genes for different conditions in yeast. Two different sets of data were used for this purpose: yeast microarray data and the data for regulatory programs in yeast. The genes from expression data were assigned to modules based on probabilistic graphical models which matched the expression of genes with the expression profiles of transcription factors such that the TF expression profile could explain the expression pattern of genes in the module [114]. The importance of identifying subnetworks over individual genes has been clearly shown in the work in which subnetworks were used to classify breast cancer metastasis. In this study, the values of gene expression from two well-known cohorts of breast cancer patients were overlaid on protein interaction networks and patterns with high discriminative metastasis potential across patients were searched. The study showed that network based classification was more accurate in predicting metastasis in breast cancer. [115]. Some other methods also used for network comparison are topological comparisons based on degree [116-119], clustering coefficient [116], path length [116], and centrality [116, 119-121], protein structure based comparisons [122-124], and dynamic Bayesian models [125].

Network comparison methods in biology have been mainly used to establish evolutionary relationships or to create phylogenetic trees from protein interaction networks. Since global network alignment is computationally intractable; most of the studies either used some heuristic method to identify an optimal alignment solution or they sought to identify smaller conserved regions in these networks (motifs, modules, or subgraphs) [102, 126-145] . Some approaches also used metabolic networks instead of protein interaction networks to establish evolutionary relationships [146-150]. Traditionally, functional annotation of unannotated proteins was based on amino acid sequence similarity alone. Currently, several homology-based approaches are being followed, of which similarity in protein networks (determined using network comparison) has become the method of choice for protein annotation [126, 151-155].

The differential subnetwork analysis work in this thesis is based on comparing both the biological as well as the topological properties of a specific molecular class of proteins (such as GFs, TFs etc.) to identify significant differential subnetworks. These properties are specific for a given condition and so we believe that our method is more robust and consistent for identifying differential components between the biological systems and for highlighting the physiological mechanisms between the different conditions. We have demonstrated this by unraveling the differential components that confer limb regeneration ability in axolotls.

Limb Regeneration

Two hundred fifty years after Lazzaro Spallanzani first demonstrated the regeneration of amputated newt limbs [156], we still do not fully understand the mechanisms of this process. The recent breakthrough of converting human adult somatic cells in vitro to embryonic stem cells has made the prospect of a regenerative medicine seem well within our grasp. Current thinking in regenerative medicine envisions the derivation, from autogeneic somatic cells, of pluripotent cells that can be directed to differentiate into transplantable replacements for cells destroyed by injury or disease [157]. Beyond this, however, is another goal, the chemical induction of regeneration directly at the site of tissue damage [158]. Achievement of this goal will require a deep understanding of the molecular components, networks and pathways that characterize regenerative competence.

Urodele Limb Regeneration

With the exception of cervid antlers [159, 160], terminal phalanges of humans and rodents [161-163], and ear tissue of certain strains of mice and rabbits, [164, 165], mammalian appendages do not regenerate after amputation. Urodele (axolotls, salamanders and newts) amphibians, which regenerate amputated limbs perfectly throughout larval and adult life, provide a research model that lends itself well to furthering our understanding of this process. Urodele limbs initiate regeneration by the formation of a blastema, a limb bud-like structure composed of undifferentiated progenitor cells. Blastema cells originate by a reverse developmental process in which the tissue matrix near the amputation plane is degraded by proteases, releasing both mature cells that are reprogrammed to a mesenchymal stem cell-like state, and muscle stem cells (satellite cells) [166-169]. The liberated cells migrate under the wound epidermis to form an avascular accumulation (also called early bud) blastema [170-172]. Once formed, the accumulation blastema is enlarged to the medium bud stage and beyond by a marked increase in mitosis [173-179]. Sustained mitosis of blastema cells, but not dedifferentiation, is dependent on factors from the wound epidermis [177] and regenerating nerves [180]. Histological [173, 174], cell marking [181, 182] and genetic marking [183] studies indicate that blastema cells derived from each tissue redifferentiate into the same tissue, although some cells derived from the dermis differentiate into cartilage as well.

Since the ability to form a blastema is what distinguishes urodele limbs from the limbs of most other tetrapod vertebrates that do not regenerate, or regenerate poorly, understanding the mechanisms that lead to blastema formation is crucial to understanding why urodele limbs regenerate, and why the limbs of other species do not. In general, the reductionist approach has been to study the individual genes or proteins involved in biological processes. With the development of high throughput technology over the last decade, there has been a shift in this approach. The ability to obtain large scale omics data has led to the development of discovery approaches that interrelate the elements of biological processes to reveal networks and pathways of organization in a system [184]. Very few studies so far have analyzed global gene or protein expression patterns during limb regeneration. In the axolotl *Ambystoma mexicanum*, expressed sequence tag (EST) resources have been developed [185] and transcription profiles of denervated vs. innervated limbs have been analyzed [186]. A number of studies have been carried out on protein synthesis and separation in regenerating urodele limbs. Autoradiographic studies of C^{14} methionine, S^{35} thioamino acids or C^{14} leucine incorporation revealed intense protein synthesis throughout regeneration [187-192]. Several protein separation analyses have been carried out using one-dimensional or two-dimensional gel electrophoresis [193-196]. These resolved up to 800 individual proteins [195] and revealed differences in protein composition at succeeding stages of regeneration in normal [194, 195] and denervated limbs [193].

Anuran Limb Regeneration

Xenopus possesses the ability to regenerate lost limbs in early tadpole stages of development, but gradually loses the capability for regeneration as development proceeds, until it is lost completely in adults [197] . Nieuwkoop-Faber (NF) stage 51–53 limb buds of the anuran *Xenopus laevis* also regenerate perfectly at any level of amputation. After NF stage 53, however, regenerative capacity becomes progressively hypomorphic and spatially restricted to progressively more distal levels, until by stage 56 or 57 amputation at any level results only in the regeneration of a muscle-less, un-segmented cartilage spike covered by an envelope of skin [198-200]. This spatiotemporal restriction of regenerative capacity is correlated with the general proximal to distal ossification of skeletal tissues, although regeneration is slightly better when amputation is through the soft tissue of the joints [201]. Loss of regenerative capacity during limb development in *Xenopus* is due to intrinsic changes in limb tissues, as shown by the fact that grafting regeneration-competent blastemas to regeneration-deficient limb stumps and *vice versa* does not alter the regenerative capacity of the blastema [202, 203]. *Xenopus* studies have focused on subtractive hybridization [204]; microarray analysis [205] and proteomics [206] for molecular screening of limb regeneration.

So we believe that regeneration-competent axolotl and -deficient *Xenopus* form excellent models to understand the differences between regeneration-competence and deficiency.

Regeneration-Competence vs. -Deficiency

The *Xenopus* and urodele limb regeneration blastema share some features. Both rely on nerve-dependent signals from the wound epidermis for their formation and growth [207-210]. Both express *prx1*, a TF that serves as an early marker of dedifferentiated cells [211, 212]. Most often, however, the *Xenopus* blastema is described as a "fibroblastema" or "pseudoblastema," as opposed to the mesenchymal nature of the urodele blastema. Although one study [213] reported that the morphology and fine structure of the cells released by histolysis is similar in amputated urodele and *Xenopus* limbs, most studies suggest that, compared to the amputated urodele limb, histolysis is limited in the amputated *Xenopus* limb, there is little if any cellular dedifferentiation, progenitor cells are fibroblastic rather than mesenchymal, muscle satellite cells do not contribute to the fibroblastema, neurovascular invasion is sparser, and the AEC is thinner with a connective tissue pad between it and the underlying cells [198, 201, 208, 214, 215]. These features have been correlated with a shift in the response to amputation brought about by the maturation of the immune system as the tadpole differentiates and undergoes metamorphosis [216-218].

Defining the cellular and molecular basis of the contrast in regenerative ability between regeneration-competent and regeneration-deficient limbs is of great interest, because of the potential to identify factors associated with successful regeneration and/or the factors that inhibit it. Differences in transcript expression by amputated regeneration-competent *Xenopus* limb buds (stage 52/53) vs. regeneration-deficient limbs (stage 57 or froglets) have been reported for specific genes and for global gene arrays compiled by subtractive hybridization or microarray [219-222]. In particular, proximo-distal axial patterning genes such as *Hoxa9, Hoxa11*, and *Hoxa13* are expressed by the fibroblastemas of *Xenopus* limbs, but their expression is not deployed in the proper spatiotemporal organization characteristic of regeneration-competent blastemas [223]. Furthermore, regeneration-deficient *Xenopus* blastemas fail to express *shh*, an important regulator of anteroposterior (AP) axial patterning in axolotl limb buds and blastemas and *Xenopus* stage 52 limb buds [222], a failure due to the epigenetic hyper-methylation of an enhancer sequence regulating *shh* expression [224]. These findings have led to the idea that faulty expression of patterning genes is the major cause of regenerative deficiency in *Xenopus* limbs [223]. The reasons why *Xenopus* limb patterning genes are not activated in their proper spatiotemporal pattern are unknown, but are likely due to an inability to activate and/or inhibit other processes necessary to the formation of a regeneration-competent blastema.

Although extensive research has been carried out to understand how the blastema is formed and which molecular entities are crucial to regeneration, very little is known about the interactive pathways and networks that lead to blastema formation in an amputated limb. In this work, we used the limb regeneration system to implement condition-specific data mining and differential subnetwork algorithms to understand the differences between the blastema formation in regeneration-competent and -deficient systems and to identify the key molecules (GFs, TFs, and ECM proteins) which might confer the regeneration ability on axolotl.

Same techniques (text mining followed by systems biology) were also used to identify the growth factors that might be used to stimulate regeneration across segmental defects. A biological screening was then done to identify a successful combination of growth factors. Of eleven growth factors identified by this method, a combination of two (BMP-4 and HGF) was shown to stimulate skeletal regeneration across 50% defects when delivered by a pig SIS scaffold. These results validated the efficacy of the methodology developed in this dissertation work for identification of the targets/key molecules in a biological system.

CHAPTER THREE: METHODS

The focus of the work developed here was to carry out an exhaustive systems level analysis to identify the differential components of a biological system. In order to achieve this task, it was crucial to integrate the vast knowledge present in the biomedical literature with high throughput experiments such as proteomics. This enabled the identification of low abundance genes/proteins which are often undetected by the high throughput experiments. This approach was applied to the limb regeneration system.

Briefly, proteomics data from limb regeneration-competent axolotl and deficient *Xenopus* was collected at different time points. Novel algorithms based on ontology matching were designed to obtain a concept list (CL) containing the relevant terms for a limb regeneration. The CL terms were used for article prioritization and classification based on the weights assigned to each article by Poisson distribution. These condition specific articles were then used to extract proteins by using exact dictionary chunking. The literature-derived proteins were integrated with the proteins obtained from the proteomics data for each condition (regeneration-competent and deficient). Protein interaction networks were hence constructed for both axolotl and *Xenopus*. These networks were used to identify molecular class based subnetworks which were compared using an innovative rule based algorithm based on both the biological and topological properties of the proteins to identify differential subnetworks between regeneration-competent and deficient systems. These differential subnetworks were also used to identify key targets of limb regeneration. Biological validation was carried out to show that key targets identified by such approach can be used to induce regeneration across a critical size defect. The following sections will describe each component of the methodology in-detail.

Although this methodology was implemented on limb regeneration, it can be used to compare any two biological systems in future. Figure 1 represents the overall methodology using breast cancer (condition 1) and colon cancer (condition 2) as an example. The goal of this pipeline then will be to discover differential components (in the form of subnetworks) between breast cancer and colon cancer. In other words, using this pipeline a researcher will be able to answer, what makes breast cancer different from colon cancer in terms of the interconnected genes/proteins?

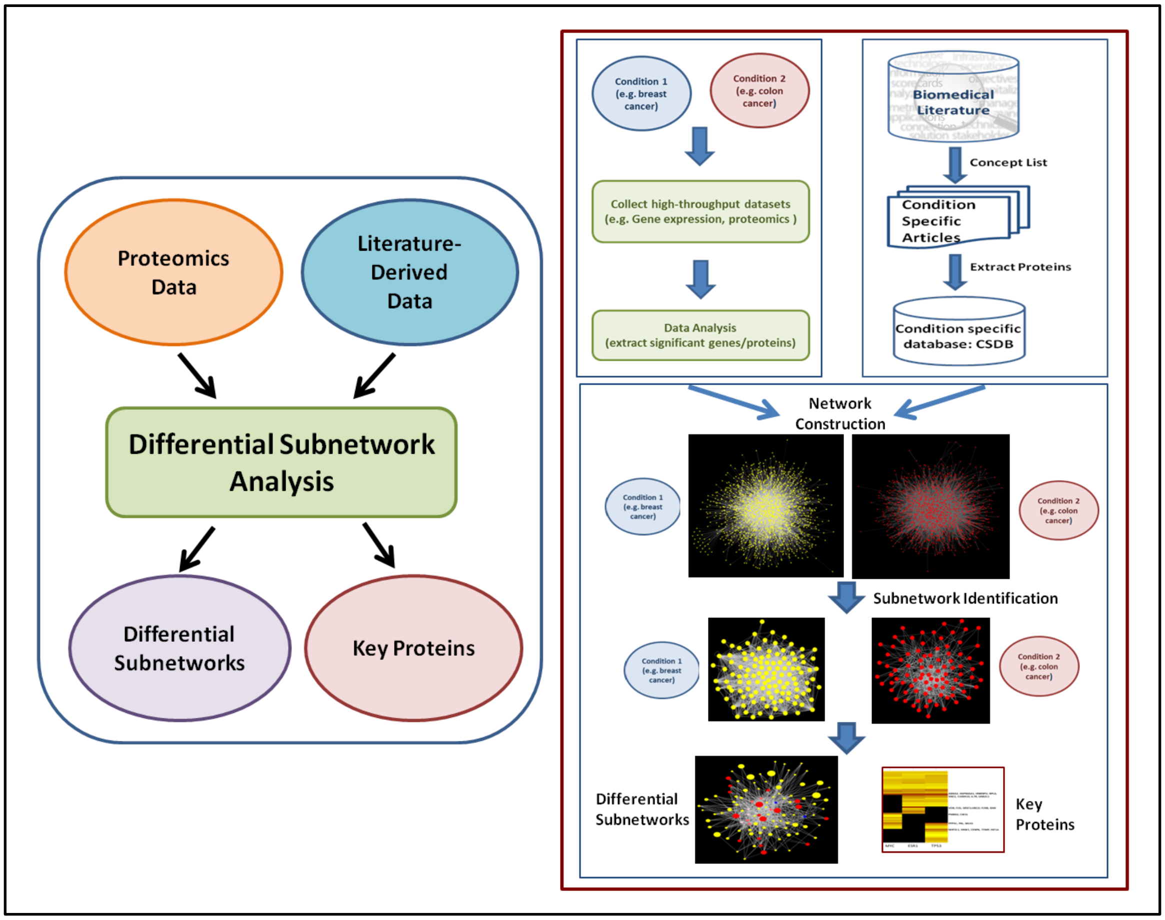
Proteomics Data
Literature-Derived Data
Differential Subnetwork Analysis
Differential Subnetworks
Key Proteins
Condition 1 (e.g. breast cancer)
Condition 2 (e.g. colon cancer)
Collect high-throughput datasets (e.g. Gene expression, proteomics)
Data Analysis (extract significant genes/proteins)
Biomedical Literature
Concept List
Condition Specific Articles
Extract Proteins
Condition specific database: CSDB
Network Construction
Subnetwork Identification
Differential Subnetworks
Key Proteins

Figure 1. Overall methodology

Proteomics

Sample Preparation

A total of 5 pools of tissue each from control, 1dpa, 4dpa and 7dpa limbs (dpa refers to days post amputation) were collected for *Ambystoma mexicanum* (will be referred to as axolotl throughout). Similar tissues were collected at 1dpa, 5dpa, 7dpa and 12dpa for *Xenopus laevis* (will be referred as *Xenopus* throughout). Each pool contained 6 tissues (from two hind limbs of three animals). The samples were processed as described earlier [225]. Briefly, flash-frozen tissues were homogenized in lysis buffer containing 8M urea and 10 mM dithiothreitol. The resulting cell lysates were denatured by urea, reduced by triethylphosphine, alkylated by iododethanol and digested by trypsin. The Bicinchoninic Protein Assay was used to determine the peptide concentration in each pool. More details about the sample collection, processing, and rest of the proteomics methodology can be found in our publications [7, 226].

Liquid Chromatography/Mass Spectrometry Analysis

Tryptic digested peptides were analyzed as previously described [225]. Samples were run on a Surveyor High Performance Liquid Chromatography (HPLC) system with a zorbax 300SB – C18 reverse column (1mmX5 cm). Each peptide pool (20 µg) was injected twice onto the column in a random order. All injections were performed using the identical equipment configuration. Peptides were eluted with a gradient from 5% - 45% acetonitrile developed over 120 min at a flow rate of 50 µl/min, and effluent was electro-sprayed into the Linear Trap Quadrupole Mass Spectrometer (Thermo-Fisher Scientific). Data were collected in the "TriplePlay" mode (Mass Spectrometry (MS) scan, Zoom scan, and MS/MS scan). The resulting data were filtered (to increase the signal-to-noise ratio) and analyzed by a proprietary algorithm developed by Higgs et al [227].

Protein Identification

Using SEQUEST and X!Tandem database search algorithms, database searches against non-redundant NCBI or International Protein Index databases were performed for peptide sequence identification. A confidence score was assigned to each peptide by q-value (false discovery rate) [227]. The score was based on a random forest recursive partition supervised learning algorithm. The % ID confidence score was calibrated so that approximately X% of the peptides with %ID confidence >X% were correctly identified [227].

Proteins were classified according to identification quality (Priority). This priority system is based on the quality of the amino acid sequence identification (Peptide ID Confidence) and whether one or more unique peptide sequences were identified (Multiple Sequences). The Peptide ID Confidence assigned a protein into 'HIGH' or 'MODERATE' categories based on the peptide with the highest peptide ID Confidence (the best peptide). Proteins with "best peptide," having a confidence between 90-100%, were assigned to the 'HIGH' category while proteins with best peptide having a confidence between 75-89% were assigned to the 'MODERATE' category. All peptides with confidence less than 75% were discarded. To increase the confidence in protein identification, the proteins were further classified based on the number of distinct amino acid sequences identified. A protein was classified as "YES" if it had at least two distinct amino acid sequences with the required ID confidence; otherwise it was classified as

"NO." Thus, the proteins with "HIGH" peptide ID confidence and with more than one identified peptide sequence were termed Priority 1. Proteins with "HIGH" peptide confidence but with only one identified peptide sequence were termed Priority 2. Priority 3 and 4 proteins were those with "MODERATE" peptide confidence with more than one and only one peptide sequence identified, respectively. Thus, Priority 1 proteins had the highest likelihood of correct identification and Priority 4 proteins the lowest likelihood of correct identification.

Protein Quantification and Statistical Analysis

Protein quantification was carried out using non-gel based and label-free proprietary protein quantification technology described previously [225, 227]. Every peptide quantified had an intensity measurement for every sample. This measurement is a relative quantity giving the area under the curve (AUC) from the extracted ion chromatogram after background noise removal. The AUC was measured at the same retention time window (1 min) for each sample after the sample chromatograms had been aligned [227]. The intensities were then transformed to the log base 2 scale (commonly used for genomic data), which served several purposes. First, relative changes in protein expression are best described by simple ratios. However ratios are difficult to model statistically, so log transformation converts ratios to fold differences. Second, the transformed data better approximate a normal distribution on a log scale [228], which is important because normality is an assumption of the Analysis of Variance (ANOVA) models used to analyze this data. Third, log base 2 is easy to understand because a 2-fold change (or doubling, or 100% increase) yielding an expression ratio of 2 is transformed to 1 (i.e. a 2-fold change is a unit change on the log base 2 scale). After log transformation, the data was then quantile normalized [229]. This normalization removed trends introduced by sample handling, sample preparation, HPLC, mass spectrometry, and possible total protein differences.

If multiple peptides had the same protein identification, their quantile normalized log base 2 intensities were weight-averaged proportionally to their relative peptide ID confidences. Then the log base 2 protein intensities were fit by a separate ANOVA statistical model for each protein. Finally, the inverse log base 2 of each sample mean was calculated to determine the fold change (FC) between samples. The maximum observed absolute FC was also given for each Priority Level. Fold Change was computed as Mean Regeneration Group/Mean Control Group. A FC of 1 means no change.

The number of proteins with significant changes for each priority was calculated. The threshold for significance was set to control the False Discovery Rate (FDR) for each two-group comparison at 5% [230]. The FDR was estimated by the q-value, as stated previously. Thus, protein fold changes with a q-value less than or equal to 0.05 were declared to be significant, leaving 5% of the determined changes assumed to be false positives.

We calculated the median percent coefficient of variance (%CV) for each priority group. Percent CV values were derived from the standard deviation divided by the mean on a % scale. The % CV was calculated for replicate variation (technical variation) and the combined replicate plus sample variation.

In constructing biological process categories, only proteins having peptide confidence levels of 90% and above and with FDR <0.05 were included. Many proteins were identified either by the same sequences or different sequences in priority 1 or 2 or both. To avoid redundancy, the fold changes of priority 1 were used if a protein was present in both the priorities, and average fold change was calculated if it belonged to the same priority. If a protein had conflicting expression patterns (upregulated in one case, but downregulated in the other) then it was not considered.

Bioinformatic Analysis

Proteins not recognized by the algorithm were manually curated. NCBI BLASTp (basic local alignment search tool for proteins) [231] was used to match the sequences of hypothetical/ novel/ unknown/ unnamed proteins against the ‘vertebrata’ category in blast (taxid: 7742) to identify their closest neighbors. Only the proteins having 90% peptide ID confidence and above and with FDR <0.05 were chosen. Accession numbers, gene names and names of the proteins were obtained from Uniprot [232] or NCBI [233] using the protein IDs obtained in the raw data. GeneCards [234] and Uniprot were used to determine their biological processes. The HPRD [235] was used to determine molecular function and primary cellular localization. EVI5 network was generated using MetaCore^{TM} analytical suite version 5.3 (GeneGo, St Joseph, MI). All non-redundant peptides having a peptide ID confidence of 90% and above were compared against EST contigs from the *Ambystoma* EST database using tBLASTn.

Condition-Specific Data Mining

An innovative methodology was developed to retrieve the condition-specific data from published literature. Condition-specific data retrieval refers to the identification of relevant or related articles from a large unknown set of articles (which can be derived from PubMed) for a given biological condition such as limb regeneration. The overview of this methodology is provided in Figure 2.

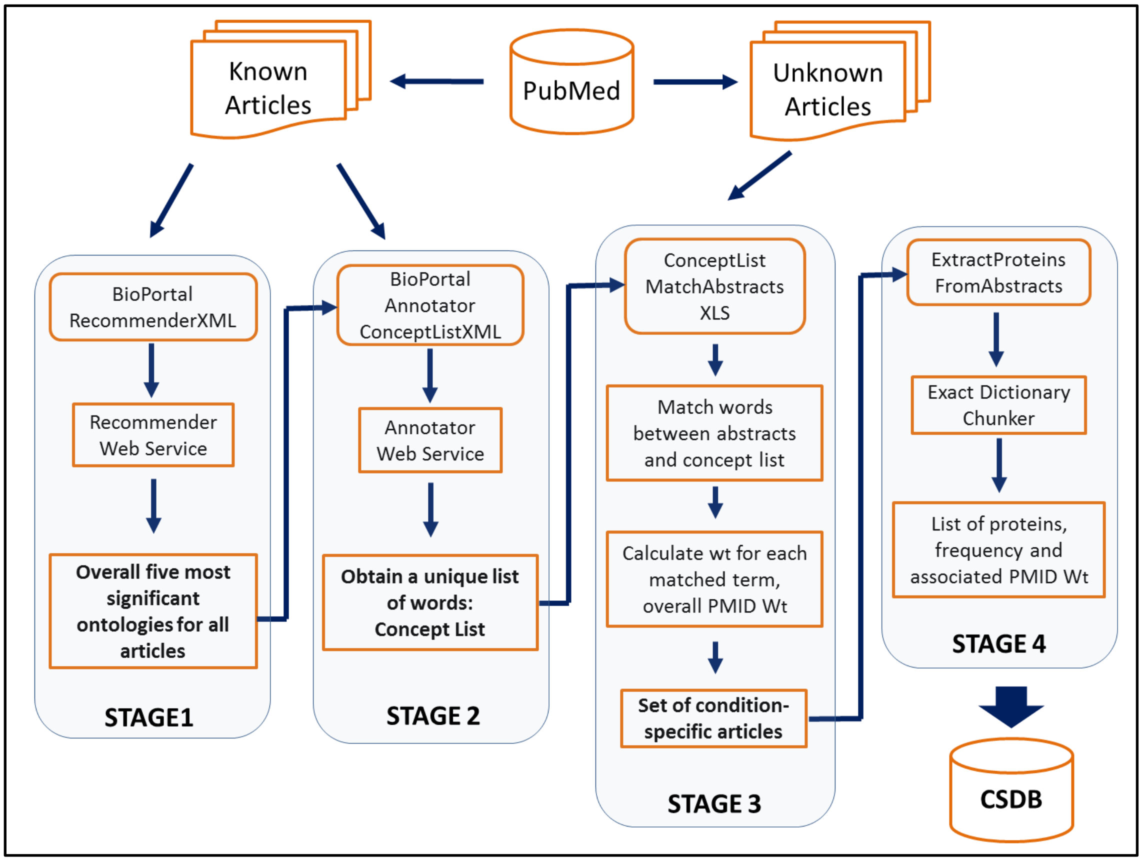
Known Articles
PubMed
Unknown Articles
BioPortal RecommenderXML
Recommender Web Service
Overall five most significant ontologies for all articles
STAGE1
BioPortal Annotator ConceptListXML
Annotator Web Service
Obtain a unique list of words: Concept List
STAGE 2
ConceptList MatchAbstracts XLS
Match words between abstracts and concept list
Calculate wt for each matched term, overall PMID Wt
Set of condition-specific articles
STAGE 3
ExtractProteins FromAbstracts
Exact Dictionary Chunker
List of proteins, frequency and associated PMID Wt
STAGE 4
CSDB

Figure 2. Overview of the condition-specific data mining methodology

In the first stage of the protocol an already known set of positive articles (for any given biological condition) is used to identify the five most significant ontologies from a collection of ontologies in NCBO BioPortal [236]. The ontologies are then used in the second stage to generate a CL. The concept list contains a collection of condition-specific terms (including but not limited to organism, tissue, cell type, biological function, and proteins studied in a given condition). The CL generated in stage two is then used as an input in stage three of the protocol for document scoring and classification of an unknown set of articles (from PubMed) to identify condition-specific articles. These articles are further processed in stage four by implementing dictionary-based methods to extract proteins which are stored in a CSDB (**c**ondition-**s**pecific **d**ata**b**ase). These stages are discussed in more detail in the following sections.

Stage 1

The Recommender web service available from NCBO BioPortal was used to programmatically retrieve the top five most significant ontologies relevant for a given biological condition. The abstracts of the articles (the known positive set for a given condition) were used as an input for the Recommender web service. This service returns a ranked list of ontologies for the text provided as an input. The ontology ranking function implemented in the Recommender uses three different scoring criteria: the number of words in the text that match with the ontologies, the number of mapping words in a ontology with other ontologies, and the total number of concepts in the ontologies [69]. One of the limitations of programmatically using this web service is that the length of each abstract (document) cannot exceed 1700 words. Hence, each document where abstract length was greater than 1700 words was split into chunks of 1700 words. Following is the pseudocode of the program used for this task.

Algorithm 3.1. BioPortalRecommenderXML

```
Input:
$A_{cs} = \{a_{cs,1}, a_{cs,2}, \dots \dots \dots, a_{cs,y}\}$ /* $A_{cs}$ is the set of known condition-specific articles in XML (eXtensible Markup Language) format, $y$ is the number of articles in $A_{cs}$ */
Output:
$O = \{O_1, O_2, \dots, O_5\}$ /* $O$ is the list of top 5 ontologies relevant to $A_{cs}$ */
Process:
Parse XML for $A_{cs}$ to extract PMIDs and Abstracts for each $a_{cs,d}$ /* $d = 1$ $to$ $y$, as defined in the input, PMID refers to PubMed identifier of an article */
for each article $a_{cs,d}$
    if length $(dl)$ > 1700 /* $dl$ is the document length */
        split the abstract into chunks of 1700 words
    end if
    Call Recommender Web Service
    Generate a list of top 5 ontologies
end for
Calculate frequency to generate the top 5 ontologies $O$ for the entire set $A_{cs}$
```

Stage 2

We defined the concept list as a collection of terms specific for a given biological condition. We hypothesized that this set can be generated by using the ontologies identified above to match the words in a set of known articles.

The Annotator web service available from NCBO BioPortal was used to programmatically retrieve the matching terms from a set of known abstracts given the top five biological ontologies identified above. The same set of abstracts as used in the Recommender was used for this step. This service matches the words/phrases in the text input (abstracts for the known documents) with the specified ontologies and returns a list of matching terms. Stop words were then removed from the matching terms. Stop words were removed using our extensive in-house list of words. These matching terms constitute the CL. However, before being used by the stage three program, the CL is further cleaned manually to remove the general words used in the biological literature (such as genes, cell). Following is the pseudocode of the program used for this task.

Algorithm 3.2. BioPortalAnnotatorConceptListXML

```
Input:
$A_{cs} = \{a_{cs,1},\ a_{cs,2}, \dots \dots \dots, a_{cs,y}\}$ /* $A_{cs}$ is the set of known condition-specific articles in XML format, $y$ is the number of articles in $A_{cs}$ */
$O = \{O_1, O_2, \dots, O_5\}$ /* $O$ is the list of top 5 ontologies obtained in Program above */
Output:
$CL = \{t_1, t_2, \dots, t_c\}$ /* $CL$ is the concept list containing terms relevant to $A_{cs}$ */
Process:
Parse XML for $A_{cs}$ to extract PMIDs and Abstracts for each $a_{cs,d}$ /* $d = 1\ to\ y$, as defined in the input */
for each article $a_{cs,d}$
    Call Annotator Web Service
    Match the words/phrases in $a_{cs,d}$ with each of the top five ontologies $O$
```

```
        Generate a list of matching words or phrases
        Remove stop words
end for
Generate unique list of terms $CL$ for the entire set $A_{cs}$
```

Stage 3

Condition-specific articles are defined as the articles from the set of PubMed articles which are relevant for a given biological condition. An innovative algorithm was designed to retrieve such articles from PubMed. This algorithm is outlined in the program below. Briefly, this code uses the abstracts and CL terms (from Stage 2) as the input. The set of articles used in the input are unknown articles or the articles for which no prior knowledge of biological condition is established. Such articles can be obtained by searching for more general keywords in PubMed. As an example, PubMed was searched for the term "regeneration." All the articles returned by PubMed were downloaded and used as an input for the program. It should be noted that although regeneration related articles were queried, it is not the biological condition of interest. The biological condition of interest is limb regeneration in this case.

The goal of this methodology was to classify regeneration articles as positive (specific for limb regeneration) or negative. It should be noted that this program can be used to run any number of articles; it can also be run on the entire PubMed set! The methodology first preprocesses the inputs by stemming and removing stop words. Second, it matches the stemmed versions of terms in abstracts with the CL terms. Third, for each matched term it generates a weight as defined by the Poisson distribution in Eq(2). The terms' weights are added to generate an overall weight or relevance score per article as described by Eq(3).

Algortihm 3.3. ConceptListMatchAbstractsXLS

Input:

$A_{uk} = \{a_{uk,1},\ a_{uk,2}, \dots \dots \dots, a_{uk,n}\}$ /* A_{uk} is the set of unknown articles in XLS format, XLS format contains PMID followed by abstract in each row, generated by parsing the XML in the program PMIDAbstractsGenerateXLSFromXMLFile, n is the total number of articles in A_{uk} */

$CL = \{t_1, t_2, .., t_c\}$ /* CL is the collection of concept list terms relevant to A_{cs}, obtained in the program above, c is the number of terms in CL */

Output:

pmid_term_weight.xls, excel sheet containing the PMID, matching term and factors used to calculate weight of each term as described by Eq(2)

overall_pmid_weight.xls, excel sheet containing the weight per PMID, as mentioned in Eq(3)

Process:

```
Set the value of $\mu = 0.013$ and $\lambda = 0.022$ /* as standardized in [58] */
for each article $a_{uk,i}$ /* $i = 1\ to\ n$, as defined in the input */
    Stem the words
    Remove stop words
    Generate CleanedAbstract
end for
for all words in $CL$,
    Stem the words
    Get unique list of stemmed words
    Add stemmed words to the list of phrases in CL to generate CleanedCL
end for
Initialize a Hashmap for CleanedCL (key = $MW_t$ , value = $V_t$) /* $MW_t$ are the matching words/phrases in $a_{uk,i}$ from CleanedCL , $V_t$ is the number of articles containing $MW_t$ , $V_t$ is set to zero in the beginning, $t = 1\ to\ c$, as defined in the input */
for the CleanedAbstract of each article $a_{uk,i}$
    Calculate the length $l$ and add to the list AbstractLength
    for each $MW_t$
        Calculate the frequency $k$ of the matching word. Add to the list Frequency
        Increment the value $V_t$ in Hashmap by 1
    end for
end for
for each $MW_t$
    calculate $idf_t = log_{10}\{n|V_t\}$ /* $idf$ is the inverse document frequency */     Eq(1)
    calculate the term weight $w_t$
    $w_t = \left(1 + \left(\frac{\mu}{\lambda}\right)^{k-1} e^{-(\mu-\lambda)l}\right)^{-1} \sqrt{idf_t}$    Eq(2)
end for
for each $a_{uk,i}$
    Calculate the relevance score $(RS)$ of $a_{uk,i}$
    $RS = \sum_{t=1}^{c} w_t$    Eq(3)
end for
```

Evaluation Metrics

The relevance score obtained by Algorithm 3.3 above was used to filter the articles and obtain condition-specific articles. The threshold for RS was decided based on the results generated by running evaluation metrics on a known set of documents 200 positive and 200 negative articles for limb regeneration. It should be noted that this set of articles was different from the set used to construct CL. Several evaluation methods mentioned were used to validate the condition-specific data mining methodology. Following are the formulas and evaluation methods used in the program EvaluationMetric.java. TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives. Mathew's Correlation Coefficient (MCC) is considered as the most standard evaluation metric for information retrieval in data mining.

$$Specificity = \frac{TN}{TN+FP} \tag*{Eq(4)}$$

$$Sensitivity/Recall = \frac{TP}{TP+FN} \tag*{Eq(5)}$$

$$Precision = \frac{TP}{TP+FP} \tag*{Eq(6)}$$

$$FScore = 2\frac{Precision \times Recall}{Precision + Recall} \tag*{Eq(7)}$$

$$Accuracy = \frac{TP+TN}{Total} \tag*{Eq(8)}$$

$$False\ Positive\ Rate = \frac{FP}{FP+TN} \tag*{Eq(9)}$$

$$True\ Positive\ Rate = \frac{TP}{TP+FN} \tag*{Eq(10)}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP)+(TP+FN)+(TN+FP)+(TN+FN)}} \tag*{Eq(11)}$$

Stage 4

The abstracts of the condition-specific articles generated above (by applying the thresholds determined by evaluation metrics) were further processed to extract proteins. This was implemented by using the exact dictionary chunker from Lingpipe [235]. The exact dictionary for proteins was created from three different sources: HPRD [21], BioGRID [22], and UniProt [20]. The official symbols, protein names and synonyms for human proteins were used in the dictionary. Three different dictionaries were used so as to include different versions of the protein names and symbols.

Condition-Specific Database

A MySQL database named condition-specific database (CSDB), was created to store data for limb regeneration. However, it can be used to store literature-derived and expression data of any other condition being investigated. Since, the data generated here is enormous and is referred by multiple programs, MySQL provided an effective way to store and query the data. Figure 3 shows the database schema with table and column names. The database contained a total of 19 tables to store information such as proteomics data, literature derived proteins, interactions, KEGG pathways (Kyoto Encyclopedia of Genes and Genomes) [237], GO biological processes [61], HPRD [21], UniProt[20], and BioGRID data [22]. A detailed description of tables and column names is provided in the Appendix (A2).

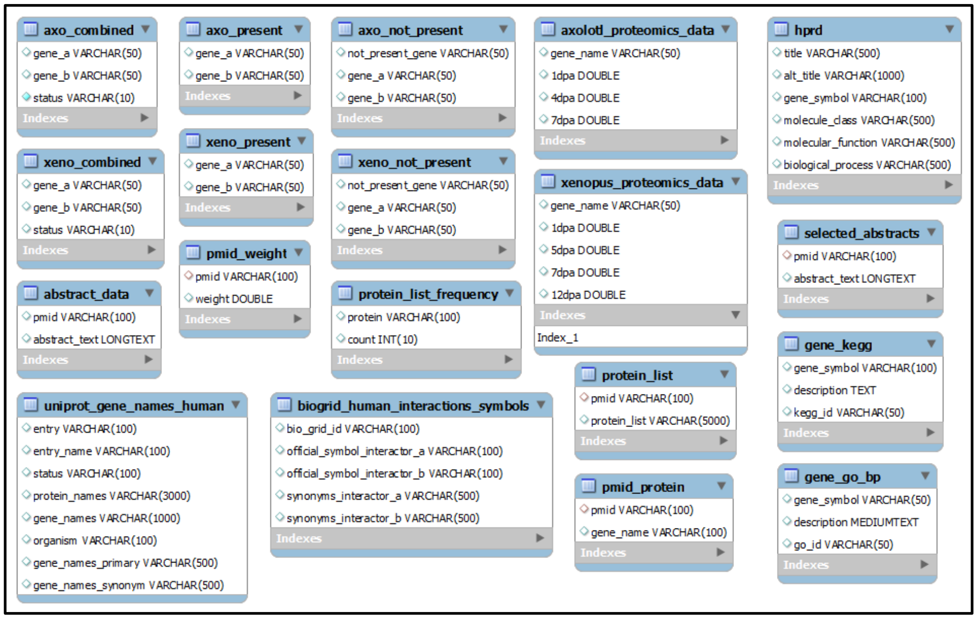

Figure 3. Condition-specific database (CSDB) schema

Differential Subnetworks

The differential subnetwork algorithm developed in this work consisted of three main parts: network construction, subnetwork identification, and differential subnetwork comparison. The following sections describe each process in detail. Figure 4 provides an overview of this methodology. Overall, in this algorithm protein interaction networks were constructed for each biological condition (two conditions to be compared). The nodes on these networks were annotated with multiple biological properties such as expression values, biological process, etc. as mentioned in the Figure 4. Subnetworks were then constructed following a rule based approach based on the end user's selection of molecular class. Multiple subnetworks were generated for each condition and all the subnetworks in one condition (for a given molecular class) were compared with all the subnetworks in the second condition. The comparison used both biological and topological properties to calculate a dissimilarity score between the subnetworks. The dissimilarity score was also used to identify the most differential nodes or proteins.

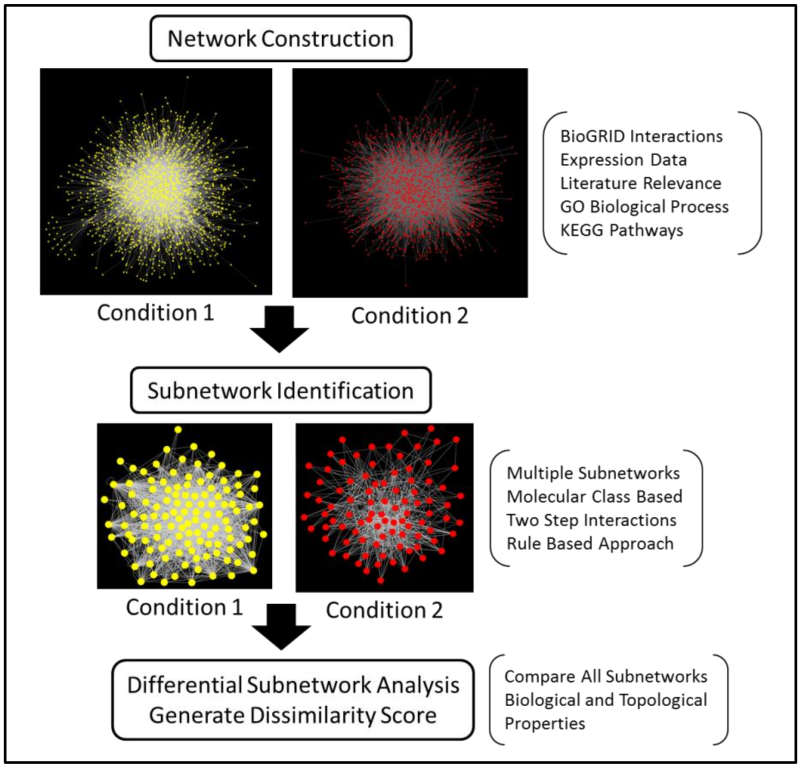

Figure 4. Differential subnetworks methodology overview

This methodology provides an exhaustive systems level comparison between two given conditions, such as normal vs. disease. The network comparison algorithm developed here considered the direction of the protein-protein interaction on the network. Most importantly, the goal of this methodology was to identify the distinguishing subnetworks and proteins across two conditions unlike the extraction of common subgraphs which has been the main emphasis of current network comparison approaches.

Network Construction

The protein interaction networks were constructed with both the proteomics and literature derived data. It should be noted that the literature derived data refers to the condition-specific data derived from the steps above. The literature and proteomics data were stored in the database tables of CSDB (refer to Appendix: A2 for details of the tables in database) which were queried in the code to generate networks. The interaction data for these proteins was obtained from BioGRID [22]. The program for network construction was implemented on limb regeneration –competent axolotl and –deficient *Xenopus*. The two networks were created with the ultimate goal of comparing the differences between competent and deficient subnetworks. The algorithm for network construction for axolotl is given below; a similar program was used to generate the network for *Xenopus*.

$G_{axo}(V_{axo}, E_{axo})$ was defined as the network for axolotl containing V_{axo} vertices or proteins and E_{axo} edges or interactions between proteins. $G_{xeno}(V_{xeno}, E_{xeno})$ was defined as the network for *Xenopus* containing V_{xeno} vertices or proteins and E_{xeno} edges or interactions between proteins. The vertices and edges were stored in the CSDB. Algoithm 3.4 is scalable and can be used for network construction of any biological condition given an expression data for that condition.

All the proteins on the network were annotated with expression values (obtained from proteomics data described above), relevance scores of literature-derived proteins (the overall weight of PMID from which a given protein was extracted as obtained by Algorithm 3.3), gene ontology biological processes [61], and KEGG pathways [238]. This information was stored in several database tables of CSDB (refer to A2 in Appendix). It should be noted that for those proteins which were associated with more than one PMID, the highest relevance score was used for annotation purposes. The program works such that expression proteins (which have the highest level of confidence since they were measured in a biological experiment) are maximized. Only those literature-derived proteins were added which connected to at least two proteins from the proteomics data. Please note that gene and protein are used interchangeably throughout the manuscript since proteins are often recognized by gene symbols.

Algorithm 3.4. NetworkConstructionAxo.java

Input:

Database tables: axo_proteomics, biogrid_human_interactions_symbols, protein_list_frequency /* Refer A2 in Appendix for the description of database tables */

Output:

Database tables: axo_present, axo_not_present, axo_combined /* Refer A2 in Appendix for detailed description */

Process:

```
for each protein in axo_proteomics
    if protein present in biogrid_human_interactions_symbols
        get the official gene symbol
        get interacting partner
        if interacting partner also present in biogrid_human_interactions_symbols
            Add to axo_present database table
             /* both the proteins are from proteomics data */
        else add to axo_not_present
            Append the gene symbol into not_present_gene column
            /* at least one protein is from proteomics data */
    else ignore that protein
end for
for each protein pair in the axo_not_present
    if the not_present_gene matches with an entry in protein_list_frequency
        Test if it is paired with at least one more protein from axo_proteomics
        Keep the protein pair in table
        /* keep only literature-derived proteins */
        /* and should pair with at least two proteins from proteomics data */
    else delete the row from table
end for
```

Subnetwork Identification

The protein interaction networks constructed for both the competent (axolotl) and deficient (*Xenopus*) systems were split into multiple smaller subnetworks based on a rule based methodology. This approach was designed to be user-centric such that subnetworks of a molecular class selected by the user were constructed. In biology, it is often the case that domain-specific researchers consider specific molecular class/es as important for a given domain. As an example, GF, TF, and ECM proteins are considered as most significant protein classes in limb regeneration. The molecular class information was obtained by parsing the downloadable XML obtained from HPRD. The program for identifying subnetworks is described below.

Algorithm 3.5. Subnetwork Identification.java

Input:

The user is asked to choose a molecular class from the list of molecular classes

Database tables: axo_combined, axo_proteomics, xeno_combined, xeno_proteomics, pmid_weight, pmid_protein, hprd, gene_go_bp, gene_kegg

Output:

For both axolotl and *Xenopus*, the following excel files are generated: interaction files for multiple subnetworks, enriched gene ontology terms and pathways, seed node files containing the seed nodes (or proteins belonging to the user specified molecular class on the respective networks) with their expression and literature relevance score.

Process:

```
for the molecular class selected by user
    identify the proteins as seed nodes in hprd which match in axo_combined
    if present in axo_proteomics
        get expression values
    else get literature relevance score from pmid_weight and pmid_protein
    for each seed node
        get the direct interactions from axo_combined
        get the second level interactions from axo_combined
        /* second level – interactions of direct interactors of seed nodes */
        /* can also be referred as two step path in the network */
        get interactions within the second level nodes
            if second level nodes connect to greater than 2 nodes
            Keep them
            else discard the interactions
            /* prevents expansion of less connected nodes */
        get biological processes from gene_go_bp
        get pathways from gene_kegg
    end for
end for
Repeat the process for Xenopus
```

Subnetwork Validation

To evaluate the significance of subnetworks, hypergeometric p-values were calculated for each subnetwork.

$$p - value = \frac{R!n!(N-R)!(N-n)!}{N!} \sum_{i\, =\, max(r,R+n-N)}^{min(n,R)} \frac{1}{i!(R-i)!(n-i)!(N-R-n+i)!} \qquad \text{Eq(12)}$$

Where, *N* = total number of unique proteins in the interaction data obtained from BioGRID, *R* = total number of proteins from the proteomics data in the axolotl network, *n* = total number of proteins in the axolotl subnetwork, and *r* = number of proteins from proteomics data in the axolotl subnetwork. Null hypothesis for the p-value was that there is no enrichment of expression proteins on the subnetworks. Expression proteins were used as a benchmark to evaluate the significance of the subnetworks since they have been biologically validated and we hypothesized that subnetworks enriched for such proteins should be significant.

Hypergeometric p-values were also calculated for establishing the significance of enrichment in biological processes and pathways for each network (*N* = total number of genes that have associated biological processes/pathways, *R* = total number of genes that have associated biological processes/pathways on the axolotl subnetwork, *n* = total number of genes for a given biological process/pathway, and *r* = total number of genes for a given biological process/pathway in the subnetwork). A description of programs written for p-value calculation can be found in the Appendix (A1). The p-values for the *Xenopus* subnetworks were calculated in a similar fashion.

Differential Subnetwork Analysis

The differential subnetwork algorithm used the subnetworks obtained in Algorithm 3.5 above to compare each subnetwork from one con dition with all the subnetworks from the other condition for a given molecular class. Common nodes were first identified between the two conditions (or subnetworks from two conditions) and these were evaluated for similarity in expression (if applicable). The algorithm then used the direct neighborhood of the common nodes to evaluate differences in both the biological and topological properties (interaction, biological processes, and pathways). A dissimilarity score (DS) was then assigned to each common node based on the differences in the properties described above. This code generated an output file, Disco.xls, containing the DS for each common node in each subnetwork comparison.

Algorithm 3.6. DifferentialSubnetworks.java

Input:

Subnetworks for axolotl and *Xenopus*: $g_{axo}(v_{axo}, e_{axo})$ and $g_{xeno}(v_{xeno}, e_{xeno})$
Database tables: interaction files for multiple subnetworks, gene ontology terms and pathways associated with each protein on the network (for a given molecular class of both axolotl and *Xenopus*)

Output:

Dissimilarity **sco**re file: Disco.xls

Process:

```
for each subnetwork in axolotl $g_{axo}(v_{axo}, e_{axo}) \in G_{axo}(V_{axo}, E_{axo})$
    for each subnetwork in Xenopus $g_{xeno}(v_{xeno}, e_{xeno}) \in G_{xeno}(V_{xeno}, E_{xeno})$
        find common nodes such that $u \in v_{axo}$ and $u' \in v_{xeno}$
        for each common node $(u, u')$
            calculate the similarity in the pattern of expression
            /* if $(u, u')$ are up/downregulated at all the time points */
            construct the k^th neighborhood of of $u$ in $g_{axo}$, $N^k_{g_{axo}}(u)$
            and the k^th neighborhood of $u'$ in $g_{xeno}$, $N^k_{g_{xeno}}(u')$ /* k =1 */
            calculate the number of shared nodes in $N^k_{g_{axo}}(u)$ and $N^k_{g_{xeno}}(u')$
            calculate the similarity in BP in $N^k_{g_{axo}}(u)$ and $N^k_{g_{xeno}}(u')$
            calculate the similarity in pathways in $N^{k}_{g_{axo}}(u)$ and $N^{k}_{g_{xeno}}(u')$
            /* Similarity for every factor $SM_f$ above is calculated as:
            $SM_f = \sum_{nd=0}^{m} \frac{C_{nd}}{m}, C_{nd}\ is\ the\ number\ of\ common\ nodes$ */
            Calculate $DS_f = (1 - SM)$ for each of the steps above
            /* $DS_f$ is the Dissimilarity for every factor */
            calculate overall dissimilarity, $DS = \sum_{f=o}^{p} DS_f$
            /* $p$ in the equation refers to number of factors with a value,
            missing values are not included in calculations */
        end for
    end for
end for
```

Programming Specifications and Visualization Software

JAVA was used as the programming language for the all the programs described in this dissertation (Java development kit version 7). The programs used open source .jar files. A complete description of all the programs is mentioned in the Appendix (A1), including the ones for which algorithms are not described in the Methodology section. All the codes were tested and deployed on a computer with 4GB memory.

The network visualizations were done with Cytoscape [78] and Circos [239]. R programming language [240] and gplots software in R [241] were used to construct the ROC curve and hierarchical clustering images for the differential subnetworks. Cluster 3.0 [91] and Java Treeview software [242] available from Stanford University were used to make the heatmaps described in the Conclusion section.

Biological Validation

Immunostaining and Image Analysis

For validation of LC/MS/MS data, immunostaining was carried out for control and regenerating limb tissues collected at 1 and 7dpa in axolotl and 5 and 12 dpa in *Xenopus*. The samples were fixed overnight in 2% paraformaldehyde in 0.8X PBS (phosphate buffered saline), rinsed with 1.0X PBS and decalcified for 30 min using immunoclear decalcifying agent (Calci-Clear Rapid, National Diagnostics, Atlanta, GA). After decalcification, the samples were cryoprotected by sequential overnight incubation in 10%, 20% and 30% sucrose in 1X PBS, then embedded in a 50:50 mixture of 30% sucrose and neg 50 frozen section medium (ThermoScientific, Waltham, MA). Sections were cut at 10 μm on a Leica CM1900 cryostat (Leica, Wetzlar, Germany) and incubated in 1X PBS to remove excess embedding medium, then blocked for 30 min in a solution of 0.01% Tween-20 and 5% milk in Tris buffered saline. For axolotl validation, sections were then incubated over night with polyclonal anti-rabbit NOS1 (Biomol International LP, Plymouth Meeting, PA) at 1:70 dilution, polyclonal anti-human fibronectin (Sigma, St. Louis, MO) at 1:400 dilution or monoclonal anti-α-actinin (Sigma) at 1:200 dilution, washed with blocking solution, incubated in the appropriate secondary antibody (goat anti-mouse AF488 or goat anti rabbit AF568, Invitrogen, Carlsbad, CA) for 40 min, washed with 1X PBS and mounted with Vectashield mounting medium containing DAPI (Vector Laboratories, Burlingame, CA). The same procedure was repeated for β1 integrin, vimentin and dystroglycan in *Xenopus* sections.

Immunostained sections were observed using the 20X objective lens on a Zeiss Axiovert 200M microscope equipped with an Apotome for optical sectioning, and images were captured with an Axiocam high-resolution camera. Sections were obtained from two hindlimbs of three animals for each time point. Six images were collected for each section, from regions located at the tip of the amputated limb to just proximal to the plane of amputation and across the putative amputation plane in control sections. Mean pixel intensities were calculated for each image by sampling 20 randomly distributed regions of each image using the measurement package of the Axiovision software (Carl Zeiss Microimaging Inc, Thornwood, NY). Statistical comparisons were performed using ANOVA. A p value <0.05 was considered statistically significant.

Segment Defect Regeneration

Multiple bone morphogenetic proteins (BMPs) have been implicated in skeletal development and regeneration [243]. We used an in-house literature-mining tool, BioMap [244], to mine the literature on fracture repair, cartilage regeneration, and bone regeneration to identify GFs in addition to BMPs that might be used to stimulate regeneration across segment defects. Keywords (similar to CL terms) related to the process of cartilage differentiation were identified and submitted to BioMap. The information extracted by BioMap was normalized using the protein and gene names from the UniProt database [20].

The HPRD was then used to identify GFs and TFs from this gene/protein list [21]. These growth factors and transcription factors were used to determine the predominant pathways and networks of protein interaction in cartilage regeneration, using MetaCore^{TM} (GeneGO Inc) [13]. These were further analyzed using four topological parameters of the CytoHUBBA plugin [79] in Cytoscape [78] to select the proteins most commonly identified as significant. The topological properties evaluated were: bottleneck nodes, maximal cliques, eccentricity, and maximum connected component.

Eleven growth factors emerged from this analysis: FGF-2, PDGF-A, PDGF-B, PDGF-D, EGF, HGF, TGF-β2, TGF-β3, Follistatin, VEGF-A, and Lefty-2. Seven of these growth factors, in addition to BMPs, were commercially available: VEGF-A, HGF, FGF-2, TGF-β3, PDGF-AA, PDGF-BB, and EGF. All of these, except EGF and PDGF-BB, had an amino acid sequence homology to the corresponding *Xenopus* growth factors (the closest amphibian to the axolotl for which such data were available) of 64% or greater. EGF and PDGF-BB were eliminated from consideration because of their low homologies (41% and 39%, respectively). Six different combinations of BMP-4 and the remaining five growth factors were tested for their ability to promote regeneration across a 50% segment defect, which exceeds the critical size defect (CSD). All growth factors were purchased from PeproTech (Rocky Hill, NJ) and stock solutions of each prepared according to instructions provided by the company.

Defects of 50% were used to test growth factor combinations. GF or tissue extract-loaded scaffolds were inserted into 50% defects and the wound was closed with two sutures of #6 silk thread (Fine Science Tools, Inc., Foster City, CA). Controls consisted of limbs in which the fibular defect received no treatment. All of the fixed limbs for each time point were first imaged by X-ray (PIXARRAY 100, Bioptics, San Jose, Costa Rica) and then by microcomputed tomography (micro-CT), using a high-resolution desktop imaging system (SkyScan 1172, Allentown, PA). Fluorochrome imaging was also carried out to measure early bone regeneration in untreated 10% vs. 50% defects. More details of the methodology can be found in Chen et al [245].

CHAPTER FOUR: RESULTS

Proteomics

Axolotl Proteomics

A total of 1624 peptides were separated in the axolotl samples. Overall summary of the proteomics data is mentioned in Table 1. Two hundred and fourteen from Priority 1 and 301 peptides from Priority 2 significantly changed between the control and axolotl samples at 1dpa, 4dpa or 7dpa. The significance threshold is set to control the FDR at less than 5%. A False Discovery is a protein declared significant when it is not. The sample median %CV for the priority 1 proteins was 12.15%. The %CV is the Standard Deviation divided by the Mean on a % scale.

Protein Priority	Peptide ID Confidence	Multiple Sequences	Number of Proteins	Number Significant Changes	Max Absolute Fold Change	Median %CV replicate + sample
1	High	Yes	281	214	3.38	12.15
2	High	No	521	301	6.94	20.99
3	Low	Yes	24	13	2.55	20.71
4	Low	No	798	469	9.88	26.35
Overall			*1624*	*997*	*9.88*	*20.88*

Table 1. Summary for axolotl proteomics data

Of these one hundred thirty-eight from Priority 1 and 285 peptides from Priority 2 were found to be statistically significant. Among these 423 statistically significant peptides, 114 peptides were not analyzed further for the reasons outlined in Methods. A total of 309 proteins were analyzed for their role in biological processes. A comparison of non-redundant peptide sequences (N=405) with the axolotl EST database identified 149 perfect-match peptides (36.8%) that were 100% identical to a translated EST contig from either *A. mexicanum* or the closely related *A. tigrinum*.

Figure 5 stratifies the proteins according to biological process (BP), and molecular functions (MF). These categories were derived from HPRD. Proteins with different biological processes and molecular functions were identified in the proteomic analysis. A detailed description of their roles and possible functions in limb regeneration can be found in our publication [7] .

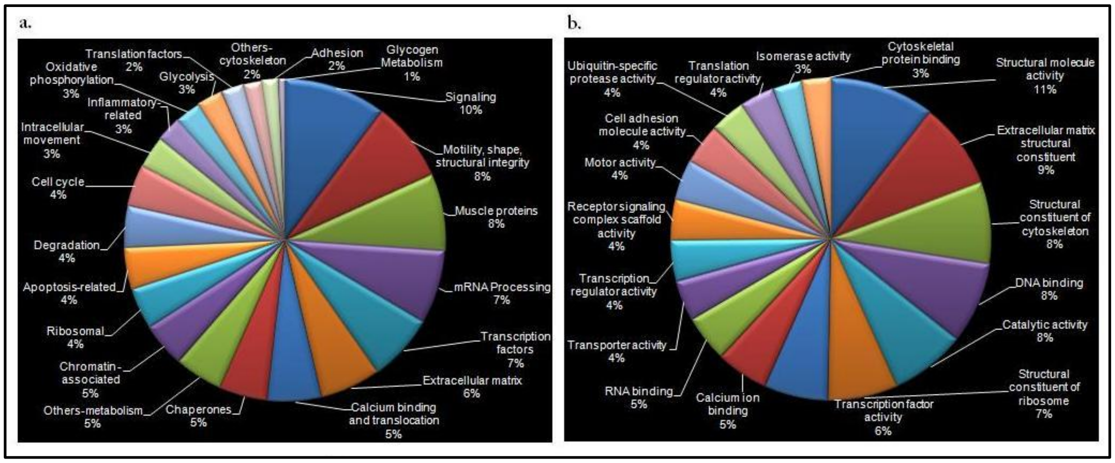

Figure 5. Biological process (a), and molecular function (b) categories for axolotl data

Xenopus Proteomics

A total of 2500 *Xenopus* peptides were separated in the samples. Table 2 provides a summary of the results. The columns in the Table 2 can be interpreted similarly to those in axolotl. There were 601 priority 1 peptides and 613 priority 2 peptides with significant change, for a total of 1214. The meaning of significant changes in proteins and %CV is the same as that described for axolotl. The sample median %CV for the priority 1 peptides was 15.97% and 31.10% for the priority 2 peptides. These were filtered as outlined in Methods to give 1014 identifiable peptides. Collapsing duplicates and discarding peptides with no known function yielded 830 proteins for analysis.

Protein Priority	Peptide ID Confidence	Multiple Sequences	Number of Proteins	Number Significant Changes	Max Absolute Fold change	Median %CV rep + sample
1	High	Yes	681	601	8.59	15.97
2	High	No	782	613	32.77	31.10
3	Moderate	Yes	94	70	11.53	28.38
4	Moderate	No	943	740	21.50	35.50
Overall			*2500*	*2024*	*32.77*	*27.88*

Table 2. Summary for *Xenopus* proteomics data

Figure 6 provides the comparison of axolotl and proteomics data and interrelationships between the function, expression, and interactions within the same data.

This figure was constructed using the Circos software [239]. The outermost circle in the figure represents the functional categories for the proteins identified by the proteomics data. The second circle is divided into proteins identified in the data under each functional category. In other words, longer length of an arc for a given function implies more proteins were identified for that function in the proteomics data. As an example, the highest number of proteins were identified for the cytoskeleton category in axolotl (represented by blue arc) and for the metabolism category (dark red arc) in *Xenopus*. The third circle highlights the proteins with greater than 2 fold change (blue dashes) and greater than 4 fold change (pink dashes). The next circles represent the fold change (green – upregulation and red – downregulation) at 1dpa, 4dpa, and 7dpa in axolotl and at 1dpa, 5dpa, 7dpa, and 12dpa in *Xenopus*. Innermost web represents the direct protein interactions among the proteins derived in proteomics analysis. More details about the functional categories of proteins can be found in our publication [226].

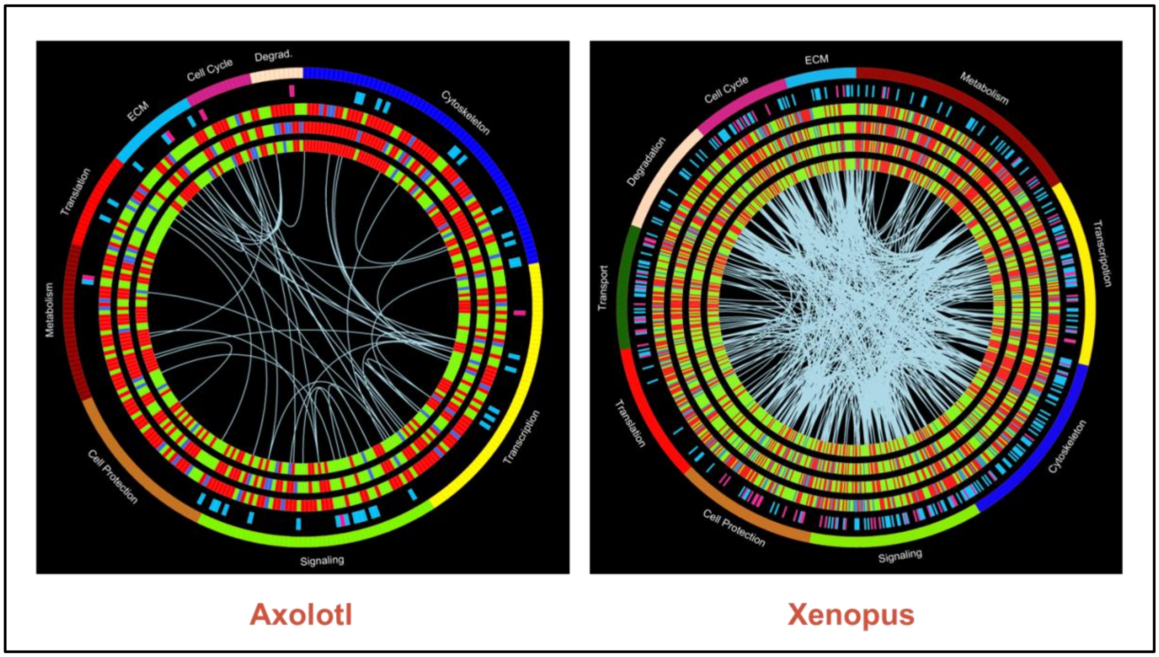

Figure 6. Circos representation of proteomics data in the axolotl and *Xenopus*

Condition-Specific Data Mining

Identification of correct and specific information from the biomedical literature is of utmost importance. PubMed is the most widely used information retrieval engine for the published biomedical literature. However, the information in PubMed is retrieved by querying for specific keywords related to the domain and is often non-specific. Moreover, the amount of articles retrieved by PubMed for any given domain is huge and makes it difficult to manually identify the correct information. Although the advancement in high throughout technologies in the recent years has generated a vast amount of data, these technologies often fail to identify low abundant proteins that are biologically crucial. We hypothesized that by developing a condition-specific data mining method, low abundant proteins can be retrieved from the published biomedical literature and can be used to augment the findings of high throughput biological experiments.

In the following sections, first the results from the preliminary work that showed the significance of extracting literature-derived knowledge are discussed followed by the results from the algorithms implemented for condition-specific data mining.

Proof of Concept

To establish the significance of extracting the data from biomedical literature, we used our in-house literature-mining tool, BioMAP [244]. Four different terms related to limb regeneration were used to query BioMAP. These terms ranged from being very specific to the biological system to being unspecific in the following order: "urodele limb regeneration," "limb regeneration," "stem cell progenitor," and "regeneration." BioMAP parsed the relevant articles for each term to extract a list of proteins. The proteins in each of the four sets were then queried against the DAVID database [246] to obtain the enriched gene ontology categories. Similar gene ontology terms were found to be enriched in each set (Figure 7).

The top ten gene ontology terms were plotted for each term, proteomics data and literature mined data, and the proteomics data alone in Figure 7. Most of the top ten terms obtained were related to the processes of development and cell cycle. These processes are closely associated with the limb regeneration system. As a validation, these findings were compared to the GO terms enriched in the proteomics data of the regeneration-competent system. Six of the top ten proteomics terms (star marked in the figure) were also obtained using BioMAP derived knowledge. A list of 1000 randomly generated proteins was also prepared to compare the findings and none of the top 10 gene ontology terms was the same as in the bibliomics or proteomics data. These results showed that the use of terms specific to a domain, in a bibliomics study, can result in condition-specific data extraction and an enrichment of relevant functions which can be used to discover novel elements.

This helped formulate our hypothesis to develop condition-specific data mining systems. However, it should be noted that processing of these terms required several manual interventions to ensure that correct data was being processed by the system since traditional text mining is prone to generate spurious results. These manual interventions motivated us to create a self-sustained, efficient, and condition-specific data mining system to extract relevant information from the published literature. The results obtained from the condition-specific data mining system are discussed below.

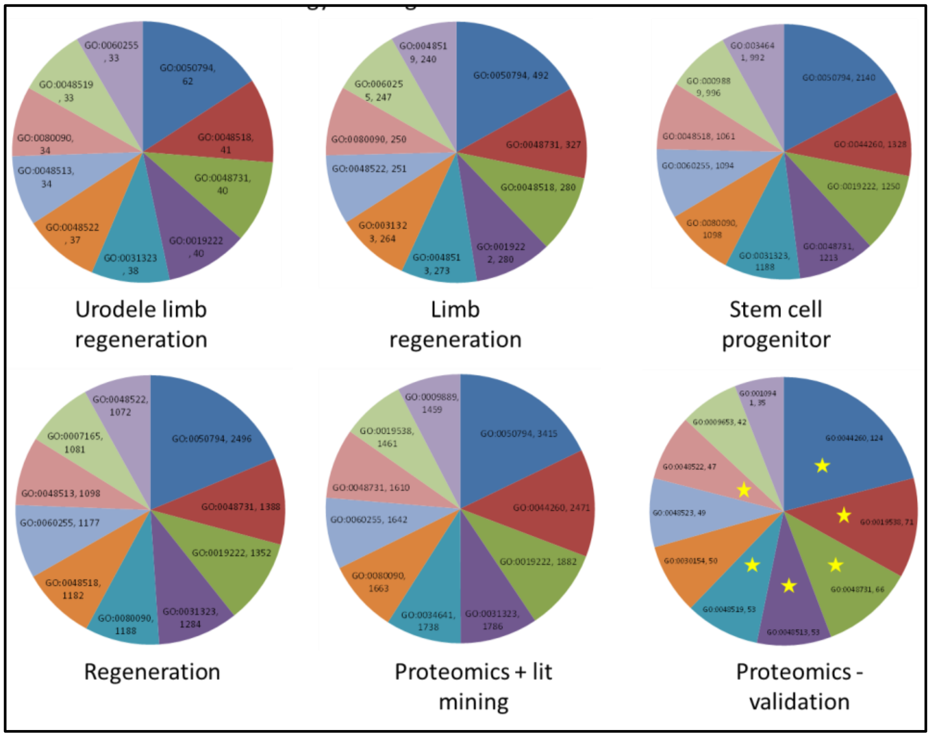

Figure 7. Significance of bibliomics knowledge

Concept List Generation

The generation of concept list was the most fundamental part of the condition-specific data mining system. The CL generated is specific for every biological condition since it requires a set of known articles from that condition to identify relevant terms. The methodology developed in this work used the biomedical ontologies made available by NCBO BioPortal [64] to create a domain-specific CL. Domain-specific refers to a biological condition of interest such as limb regeneration. The Recommender services were first used to programmatically retrieve the five most significant ontologies related to limb regeneration. For this purpose, a set of 300 articles was manually identified as a positive set of articles for limb regeneration. These articles were obtained from PubMed by querying with the keyword “limb regeneration.” Domain experts used their knowledge to identify limb regeneration specific articles from this set. Of these 300 articles, 283 were finally used to query the Recommender web service since the remaining either did not have abstracts or were written in a different language. These articles were downloaded in the XML format and Algorithm 3.1 described in the methods was used to parse the XML to extract abstracts and PMIDs.

NCIT, SNOMEDCT, NIFSTD, MESH, and EHDA were identified as the most significant ontologies for the limb regeneration domain. The National Cancer Institute Thesaurus (NCIT) contains terms relevant for clinical care, translational and basic research, and public information and administrative activities. This ontology contains 110,375 classes with 173 properties [247]. The Systematized Nomenclature of Medicine – Clinical Terms (SNOMEDCT) contains clinical terms with 3,000,542 classes. The Neuroscience Information Framework Standard Ontology (NIFSTD) comprehensively describes neuroscience data and resources. A total of 108,426 classes and 627 properties are present in this ontology. The Medical Subject Headings (MeSH/MESH) consists of terms that describe the content of an article (keywords associated with articles). A set of 245,887 classes comprise MESH. The Human Developmental Anatomy Ontology (EHDA) contains 8,340 classes describing the stage-specific human anatomical structures.

Of the ontologies identified as most significant, SNOMEDCT and MESH are common to most of the biological domain because of their wide coverage. However, NCIT, NIFSTD, and EHDA were more specific to the limb regeneration domain since a lot of terms present in limb regeneration literature matched with these ontologies. Although a manual selection of ontologies related to a biological domain is possible, such a methodology is not scalable and since ontologies are rapidly being updated, programmatic retrieval is the most efficient method. It can also be argued that less or more than five ontologies can be used, our research indicated that five ontologies were the most optimum as they cover most of the important terms in the text as determined by the domain expert. However, if for a given biological domain there is a need to change the number of retrieved ontologies, it can be easily specified in the program.

The ontologies identified by the Recommender were then used to query the Annotator web service, as described in Algorithm 3.2, with the same set of articles to identify the terms in the abstracts that match the ontologies. This collection of terms was termed as the CL for the given domain. Table 3 provides an example of terms in the CL for limb regeneration. It should be noted that the CL is comprised of terms in the second column of Table 3, the first column represents the categories of the terms. As can be noticed in Table 3, the terms represent different aspects of the biological system such as organisms used in the studies, tissues, cell types, and even the biological functions and proteins studied by researchers in the domain of limb regeneration.

Category	**Terms in Concept List**
Cell Type	Adult Stem Cells
Organism	*Xenopus laevis*
Organism Class	Urodela
Biological Function/Process	Mesenchymal Cell Proliferation
Biological Function/Process	Dedifferentiation
Tissue	Blastema
Genes/Proteins	Hox

Table 3: An example of concept list terms in limb regeneration

A total of 2,798 unique terms were present in the CL for limb regeneration. This CL contained similar terms (wound, wounds) since stemmer was not used at this step to preprocess the abstracts. The articles were preprocessed for removing stop words but stemmer was not used since the stemmed terms cannot be matched with the ontology. However, the stemmer was used later in the mining to overcome this problem.

The CL derived by using the annotator functionality had some limitations such as the presence of general science words (gene, cell, DNA) and some other English words (variable, value, symmetric) which were not present in the stop word list. These words were identified since some of the ontologies, especially SNOMEDCT contain a wide range of words. This problem can be dealt with by either getting rid of such ontologies or by adding these words to the stop word list. In our opinion, these were not good solutions because of the following two important reasons: a lot of important domain-specific terms which match with such ontologies will be lost, addition of terms to the stop word list will defeat the purpose of domain-specific mining since some of the general science terms can be important in other domains. As an example, the word DNA can be important if DNA repair is being studied. TF-IDF was also explored as a method to clean the CL but it got rid of many important words such as regeneration, amphibian. TF-IDF was not a good choice to clean this list since the CL is built from a positive set and some of the important terms are overrepresented in these articles. Hence, we preferred manual curation by the domain expert to get a filtered and more specific CL. Since, the original CL for limb regeneration (2,798 words) was not huge, manual curation was not a time consuming task. The final CL (after manual curation) consisted of 687 terms.

Identification of Condition-Specific Articles

To identify condition-specific articles for limb regeneration, a large set of unknown articles was downloaded from PubMed in the XML format. The XML format was parsed to extract the abstracts and PMIDs (refer to A1 in the Appendix for a description of the codes) into the XLS format. Before this step, the abstracts and PMIDs were stored in cache but in this step since the number of articles was huge, it was important to convert the format and use it as an input file. This greatly increased the efficiency of running the program and reduced the computational time. PubMed was queried for the keyword "regeneration" and a total of 218,249 articles were downloaded. Of these, 172,986 articles were further processed (remaining articles did not have an abstract or were written in another language and hence were ignored).

The program ConceptListMatchAbstracts.java was used to match the CL terms with the words/phrases in the articles. Each term in the CL (after stemming CL contained 652 unique terms) that was found in a given PMID was assigned a weight as described in the Methodology section (an example of the results is described in Table 4). The columns containing the values k (frequency of the word in article), l (length of the article), idf (inverse document frequency of a term) were used as factors to calculate the weight for each term (described in the Algorithm 3.3 in the Methodology section in detail).

The Poisson distribution assigned term weights such that the terms specific for limb regeneration were given higher term weights (such as "stump" from Table 4 has $w_t = 0.87$) as compared to terms which were more general (as an example, "regeneration" from Table 4 has $w_t = 0.24$). All the term weights were averaged to assign an overall weight or relevance score (RS) for a given PMID (Table 5 provides an example). Therefore, if a given PMID has terms that are more specific then it is assigned a higher RS.

PMID	Word/Phrase from Concept List	Stemmed Word	k	l	idf	Weight (w_t)
1292570	regeneration	regener	3	154	0.32	0.24
1292570	limb, limbs	limb	4	154	0.33	0.31
1292570	amputation	amput	1	154	0.61	0.16
1292570	growth	growth	2	154	0.74	0.26
1292570	amphibian, amphibians	amphibian	1	154	0.82	0.18
1292570	proliferating, proliferation	prolifer	3	154	1.10	0.44
1292570	developmental	development	1	154	1.32	0.23
1292570	cell proliferation	cell proliferation	2	154	1.56	0.37
1292570	leg, legs	leg	1	154	1.90	0.28
1319553	regeneration	regener	1	97	0.32	0.17
1319553	amputation	amput	2	97	0.61	0.32
1319553	newt, newts	newt	2	97	0.64	0.33
1319553	forelimb, forelimbs	forelimb	2	97	0.98	0.41
1319553	stump	stump	6	97	1.03	0.87
1319553	denervation	denerv	4	97	1.19	0.73
1319553	dedifferentiation	dedifferenti	1	97	1.28	0.33
1319553	innervation	innerv	3	97	1.32	0.63

Table 4. Output file example from ConceptListMatchAbstracts.java

PMID	Relevance Score
3665773	10.95
7813787	10.19
8877452	9.98
6474177	9.67
9389454	9.12
3698099	9.05
8150219	9.02
9527876	8.97
1569412	8.59
3183582	8.45
2471654	8.14
2005423	7.93
3464959	7.80
2092016	7.65
2552324	7.63

Table 5. Overall PMID weight

Poisson distribution is a very well established and effective method for estimating term frequencies in information retrieval. Two of the previous methods such as the bm25 and PMRA model (as described in the Background section) are known to outperform the other methods for information retrieval in the biomedical domain. Both these methods use similar Poisson distributions to estimate term frequency. However, a few things set our work apart from these already existing methods. First, CL was used to gather information from several different ontologies and is specific to a given biological condition such as limb regeneration in this case. The methodology described here starts with a collection of CL terms from a known set of articles and hence is more specific to a given condition. The existing methods on the other hand utilize general keywords (query term, MeSH, article title) to generate a list of related articles. Often times, these articles are not specific to the domain. Secondly, the related articles displayed in PubMed (uses the PMRA model) are a very small number and cannot be used in an automated way to extract information such as protein names. There are some other methods described in the Background section which allow automated retrieval of articles for a given condition. However, these methods are often limited to approximately 10,000 related articles for a given domain. The method developed here can be used to query a much larger number of domain-specific articles.

The articles were classified as positive for limb regeneration by setting a threshold value of 2.5 for the overall PMID weight (RS). A total of 64,417 articles were classified as limb regeneration specific articles from a set of 172,986 articles. The range of overall PMID weight (or RS) for the entire set of 172,986 articles varied from zero (no terms for limb regeneration in a given PMID) to 14.96. Figure 8 depicts the graph for the number of articles or PMIDs in a given RS range.

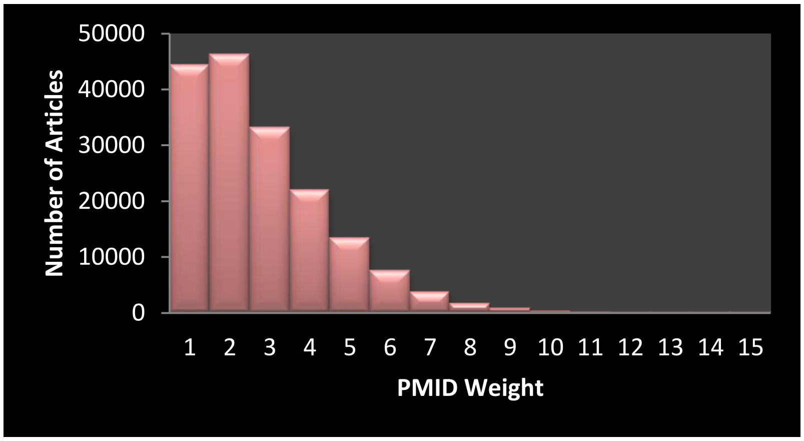

Figure 8. PMID weight vs. number of articles

The number on the x-axis in Figure 8 represents the upper value of the range for every bar. It is evident from the figure that most of the articles in the set were in the range of zero to two (90,529) which were negative (unspecific) for limb regeneration. Note that the articles in the range 2-2.5 were also classified as negative (18,040). There were 488 articles with a score greater than or equal to 10, and only 9 articles with a score of 14 or above. This showed that very few articles contained a very high number of terms from the CL.

Data Validation

The results generated above were validated by testing this methodology on a set of 200 positive (different from the set used to construct CL) and 200 negative articles for limb regeneration. The PMID weights obtained for this set of 400 articles were evaluated by different metrics: sensitivity, specificity, precision, f-score, accuracy, and Mathew's Correlation Coefficient (MCC). In the field of IR, MCC is considered as the most standard measure to evaluate the validity of the data. The best MCC value of 0.916 was found for the PMID cutoff weight of 2.5 (highlighted in Table 6 and Table 7). Hence, this was used as a threshold for classification of condition-specific articles for limb regeneration. A very high MCC value also proves the validity of the model. Table 6 shows the values for the evaluation metrics.

Threshold	Specificity	Sensitivity	Precision	F-Score	Accuracy	MCC
0.25	0.50	1.00	0.67	0.80	0.75	0.58
0.5	0.64	1.00	0.73	0.85	0.82	0.68
0.75	0.75	0.99	0.80	0.88	0.87	0.76
1.0	0.80	0.99	0.83	0.90	0.89	0.79
1.25	0.86	0.98	0.88	0.92	0.92	0.85
1.5	0.89	0.98	0.89	0.94	0.93	0.87
1.75	0.93	0.98	0.93	0.96	0.96	0.91
2.0	0.94	0.97	0.94	0.95	0.95	0.91
2.25	0.97	0.95	0.96	0.96	0.96	0.92
2.5	0.98	0.94	0.97	0.96	0.96	0.92
2.75	0.98	0.92	0.98	0.95	0.95	0.90
3.0	0.98	0.90	0.98	0.93	0.94	0.88
3.25	0.99	0.87	0.98	0.92	0.93	0.86
3.5	0.99	0.82	0.99	0.90	0.91	0.82
3.75	0.99	0.78	0.99	0.87	0.89	0.79
4.0	1.00	0.74	0.99	0.84	0.87	0.76

Table 6. Evaluation metric results

Table 7 depicts the True positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) values for the different thresholds. The values for the threshold of 2.5 are highlighted in Table 7.

Threshold	TP	FP	TN	FN
0.25	200	100	100	0
0.5	199	72	128	1
0.75	198	51	149	2
1.0	197	41	159	3
1.25	196	28	172	4
1.5	196	23	177	4
1.75	196	14	186	4
2.0	193	12	188	7
2.25	190	7	193	10
2.5	188	5	195	12
2.75	184	4	196	16
3.0	179	4	196	21
3.25	174	3	197	26
3.5	164	2	198	36
3.75	156	2	198	44
4	147	1	199	53

Table 7. Contingency values for evaluation metrics

A Receiver Operating Characteristic (ROC) curve is often used to visualize the performance of a binary classifier by plotting the false positive rate (FPR) on the x-axis and the true positive rate (TPR) on the y-axis. This data is plotted for FPR and TPR values at different thresholds. Figure 9 depicts the ROC curve for this data with the interval of 0.25 from the lower range to higher range of the PMID weights.

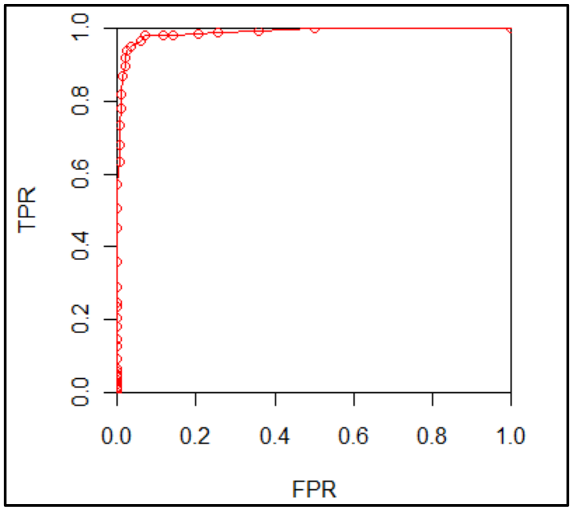

Figure 9. ROC curve for the validation data

Protein Extraction

Once the condition-specific articles were identified (64,417 articles greater than overall PMID weight of 2.5), they were further processed with the program ExtractProteinsFromAbstracts.java. The program used the exact dictionary chunker available from lingpipe [235]. The dictionary created for this purpose was derived from the protein names, official gene symbols and their synonyms from three different resources: HPRD, BioGRID, and UniProt. These resources to our knowledge provide the most comprehensive coverage of proteins. The gene symbols and full names of proteins were processed only if they exactly matched the dictionary, partly matching names were not used to avoid discrepancies. The official gene symbol of all the proteins and the PMIDs (and the associated weight from condition-specific data mining) from which the proteins were derived were stored in the condition-specific database (refer to A2 in the Appendix for a description of the database tables). Of the 64, 417 condition-specific articles, 31,751 articles were identified as containing one or more protein. A total of 5,273 unique proteins were extracted from these articles. We will refer to this set as the literature-derived limb regeneration proteins.

Network Analysis

The major goal of any bioinformatics study has been to compare the disease (or any other biological condition) sample with the normal sample, mainly on the basis of differential gene expression, so as to identify significant genes associated with the disease phenotype. However as mentioned before, such approaches suffer from several limitations. The subsections below first highlight our previously published work which helped establish the significance of network analysis. The later subsections describe the detailed results of the differential subnetwork analysis approach as implemented in this dissertation.

Proof of Concept

In our previous work, the proteomics data of regenerating axolotl limbs was analyzed using the commercial tool, MetaCore^{TM}. The expression of proteins at three different time points, 1, 4 and 7 dpa was monitored and several significant proteins were identified based on their differential expression. We particularly focused on a protein, EVI5, and proposed its importance in the process of limb regeneration based on its interacting partners. The interaction profile of EVI5 confirmed its role in the process of cell cycle that is very critical for regeneration to occur [7]. Figure 10 shows the network of EVI5.

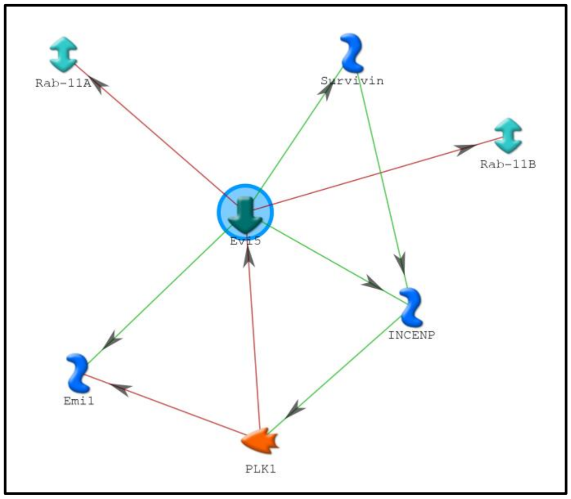

Figure 10. EVI5 network

In another study on axolotl limb regeneration, TFs that might be responsible for regulating the process of regeneration in axolotl limbs were identified. The main emphasis of this study was to show the importance of adding literature knowledge along with the high throughput data. The five most connected factors, c-Myc, SP1, HNF4A, ESR1 and p53 were found to regulate ~50% of the proteins from proteomics data. Among these, c-Myc and SP1 regulated 36.2% of the proteins. c-Myc was the most highly connected TF (71 targets). Figure 11 shows the network of these TFs. All the green colored circles on the network were nodes or proteins from the proteomics data. The circles representing the TFs were sized based on their connectivity to the proteomics proteins (larger size implies higher connections) and connections were visualized by white lines connecting the TFs with the proteomics proteins. Transcription factor network analysis showed that TGF-β1 and fibronectin (FN) lead to the activation of these TFs. We also found that other TFs known to be involved in epigenetic reprogramming, such as Klf4, Oct4, and Lin28 were also connected to c-Myc and SP1. Figure 12 highlights the connections between these proteins. The TFs identified here were not present in the proteomics data but were found to be connected to several differentially expressed proteins. This demonstrated the advantage of incorporating the bibliomics data. In this study, a possible link between stem cell factors and the proposed TFs for limb regeneration was also established [6].

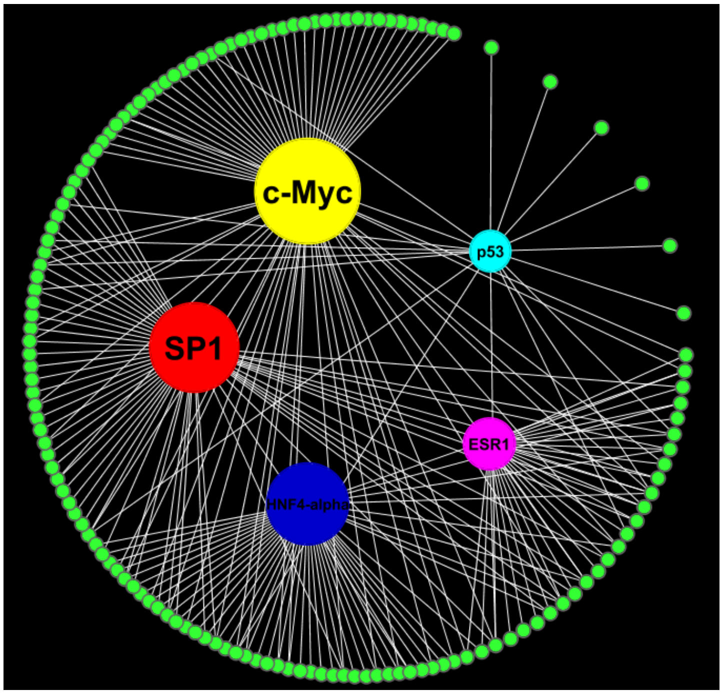

Figure 11. Transcription factor network with the proteomics derived data

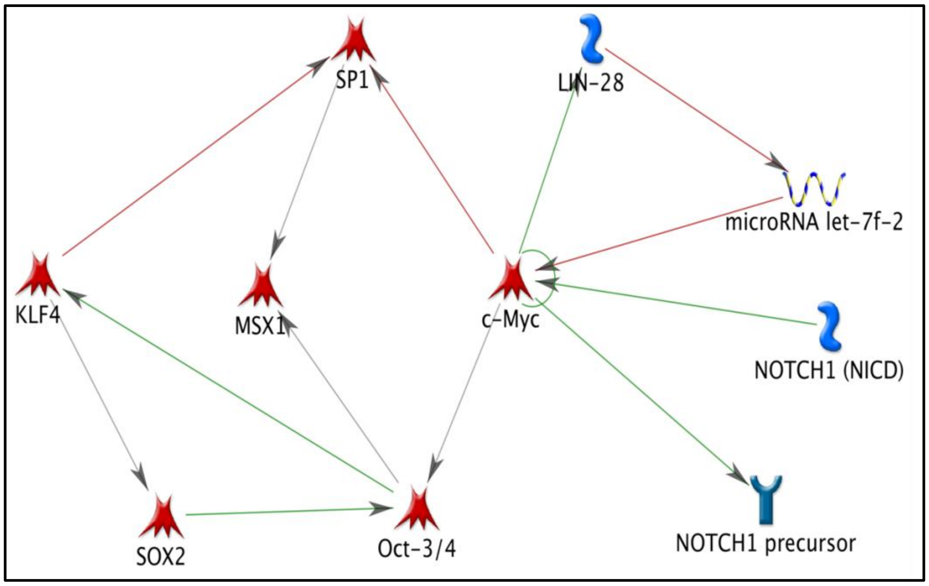

Figure 12. Transcription factor and stem cell factor network in limb regeneration

A small network comparison study was also carried out to establish proof of concept for differential subnetwork analysis. It was implemented to identify the differences between limb regeneration-competent and –deficient model systems. The proteomics data was used in combination with the literature data to draw these two networks (Figure 13). Networks were drawn using the visualization software, Cytoscape [78, 248].

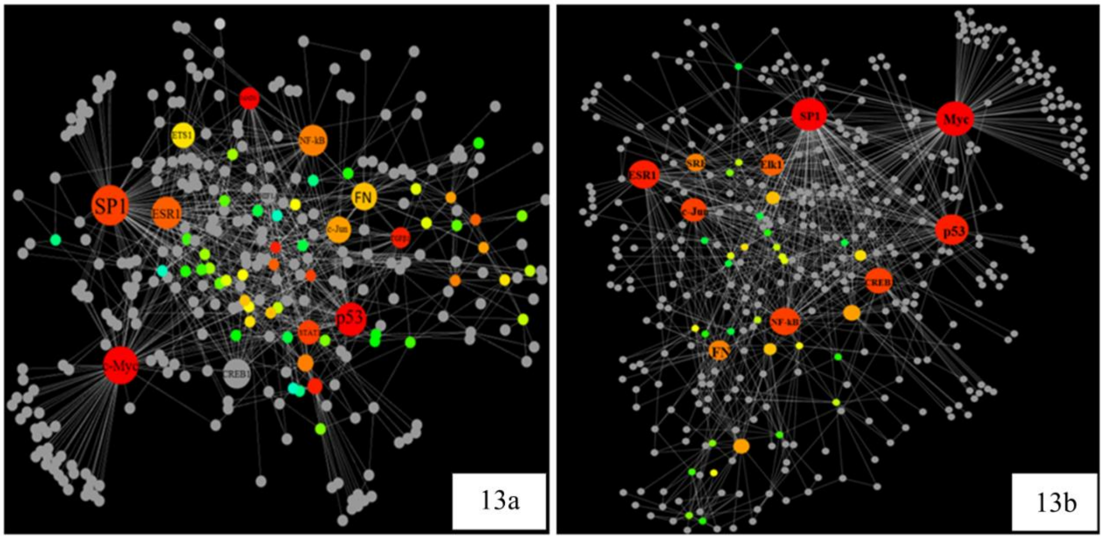

Figure 13. Networks of Axolotl (13a) and *Xenopus* (13b)

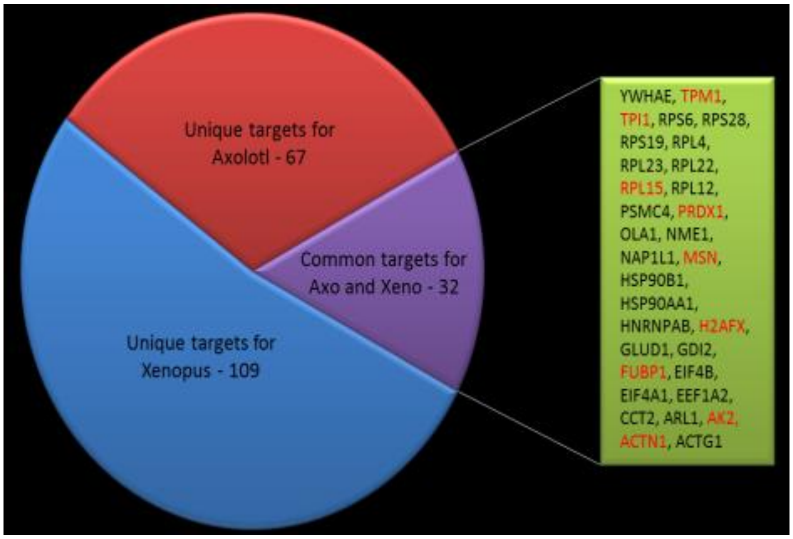

Figure 14. Targets of c-Myc in the proteomics data

Interestingly, it was found that even though many important nodes were same in both the networks, their interacting partners were vastly different. As an example, for c-Myc 67 and 109 unique targets were found in axolotl and *Xenopus* networks respectively. Only 32 targets were common to both the networks and these targets could be further differentiated with respect to expression (Figure 14). Only nine (marked in red in Figure 14) common targets showed a similar expression in both the systems. We defined similar expression as either up or downregulation. This indicated a major change in the connectivity of the same protein in the deficient system leading to completely different biological processes. This further lead us to believe that network comparisons can reveal the underlying patterns which distinguish biological systems. Note that this comparison was manually performed to understand the difference in the interaction profile of important nodes between the regeneration-competent and –deficient systems.

To understand the significance of assigning a biologically relevant score to the proteins, a pilot experiment with breast cancer data was carried out. Differentially expressed, 100 random genes were selected from one of the hallmark studies in breast cancer [249]. A protein interaction network was constructed by overlaying the expression data of these genes. The proteins identified from the literature which connected to expression data proteins were also included in the network. Cancer specific pathway information from NetPath (an in-built feature of HPRD) [21], expression information, functions, and degree (topological parameter) were used as parameters for the biological scoring of the nodes. For topological scoring, only the degree parameter was used. These methods were compared for the identification of the top 10 nodes. Among the top 10 nodes (colored red in Figure 15, the size of the nodes reflects significance, larger size entails significance), only two nodes were common to these scoring methods. A further analysis showed that biological scoring (Figure 15a) identified nodes other than hub nodes while topological analysis (Figure 15b) identified mainly hub nodes on the network. Four of the top 10 nodes identified by biological scoring belonged to the Wnt pathway, known to play a very important role in breast cancer [250]. This helped establish the significance of using biological knowledge along with topological knowledge.

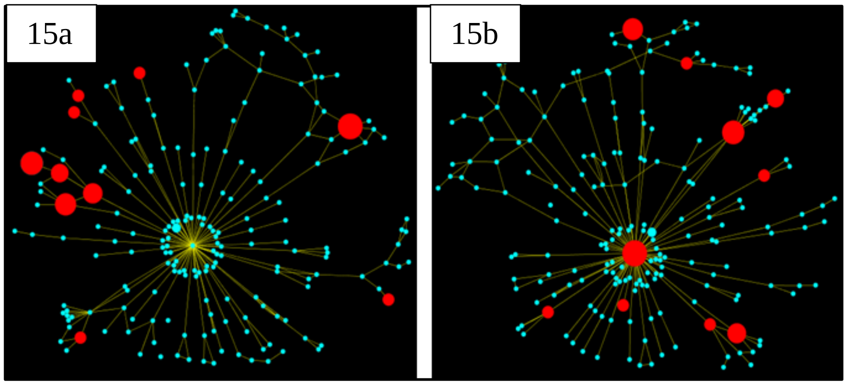

Figure 15. Biological scoring vs. topological scoring for breast cancer proteins

The work discussed so far in network analysis helped establish the significance of network analysis and the use of biological and topological properties for scoring of the nodes. However, as emphasized earlier most of this work was done manually or using commercial software. The commercial software did not have the ability to do differential subnetwork analysis. This motivated us to develop an approach for differential subnetwork analysis and the results from this approach are discussed below. The differential subnetwork analysis involved three main steps: network construction, subnetwork identification, and differential subnetwork analysis. The results from each of this step are discussed below.

Condition-Specific Network Construction

The goal of network construction was to create protein interaction networks with both proteomics and literature derived proteins such that the proteomics (or any expression data) proteins are maximized on the network. Two networks, one each for axolotl (limb regeneration competent network) and *Xenopus* (limb regeneration deficient network) were constructed using the Algorithm 3.4. As described in condition-specific data mining, a total of 5,273 unique proteins were obtained by mining articles specific for limb regeneration.

The proteomics data of axolotl contained 309 proteins of which only 263 had associated interactions reported in the BioGRID data and so the rest of the proteins were not included in any further analysis. To construct the axolotl network, interactions for the 263 proteins from the proteomics data were obtained by using BioGRID [22]. As mentioned in the proteomics Methodology section, the axolotl and *Xenopus* proteins were converted to human orthologs. So, the interactions in the BioGRID database were filtered for *Homo sapiens* before using them for the purpose of network construction. Of two hundred sixty three proteins from the proteomics data, 168 proteins had 493 direct interactions among themselves (as obtained from BioGRID).

The literature-derived limb regeneration specific proteins were added to the network only if they had direct connections with at least two proteins from the proteomics data. Hence, the networks were enriched for proteins from the expression data. This ensured that only proteins important in limb regeneration from the literature were added and so provided more validity to the methodology. Any random protein that could have been extracted by mining (perhaps as the result of false positive detection) should have been eliminated at this step since it is highly unlikely that a random protein was identified in condition-specific data mining and also connects two proteins from the proteomics data. By using this methodology only 984 literature-derived proteins (of 5,273) with 5,826 interactions were added to the axolotl data. Overall, the axolotl network consisted of 1,244 proteins (nodes) and 6,319 interactions (edges).

The *Xenopus* network was constructed in a similar fashion to the axolotl network. Eight hundred thirty proteins were present in the *Xenopus* proteomics data of which only six hundred and one proteins had associated interactions in the BioGRID database. It consisted of 1,634 literature-derived proteins that had 13,745 interactions of with the proteomics data proteins. Overall, the *Xenopus* subnetwork contained 2,235 nodes with 16, 582 interactions.

The proteins on the protein interaction networks were further annotated with the following feature vectors: expression, literature relevance, biological processes, and pathways. Figure 16 below shows the condition-specific networks constructed for axolotl and *Xenopus.* In these networks, pink and blue nodes were derived from literature while yellow and red nodes were derived from proteomics data. Overall, 1017 nodes were common to both axolotl and *Xenopus* networks; 227 nodes were unique to axolotl network and 1218 nodes were unique to the *Xenopus* network.

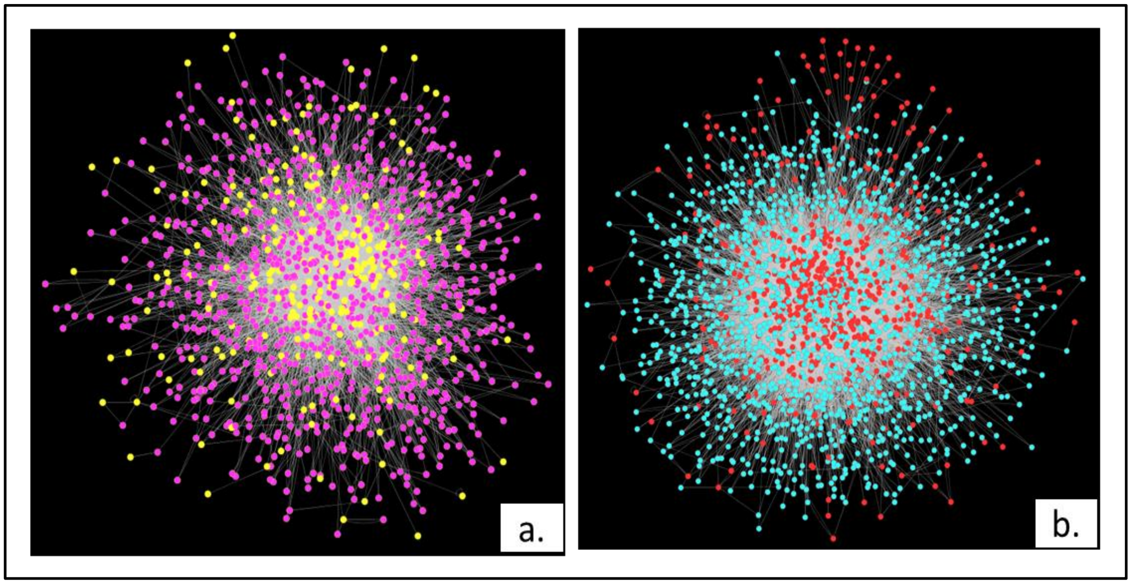

Figure 16. Condition-specific networks in axolotl (a) and *Xenopus* (b)

Subnetwork Identification

Subnetworks provide more consistent representation of the functional components of a biological system. Subnetworks (as mentioned in the Background section) are also known to have better prediction power in disease prognosis as opposed to individual genes [115]. However, most of the time domain-specific research needs are different and specific molecular classes of proteins are of interest to the biologist. As an example, in certain disorders such as cancer, the molecular class “kinase” is known to play a major role and several kinases are being studied for their potential use as drug targets. On the other hand, in the field of limb regeneration, immense importance is given to GFs, TFs, and ECM proteins.

We developed a user-centric molecular class based system for subnetwork identification (Algorithm 3.5). The user can specify the molecular class of interest and the proteins belonging to that class are then searched on the network. These proteins are referred as seed nodes that are used to build subnetworks. Each of the subnetworks is also functionally annotated with p-values for enriched biological processes and pathways. These subnetworks can then be compared between two biological conditions to identify differential subnetworks.

Growth factor, transcription factor, and extracellular matrix protein subnetworks were identified for both the axolotl and *Xenopus* systems. Table 8 below shows the overall summary of GF subnetworks. The second column in the Table 8 shows the GF subnetworks in axolotl (A) and *Xenopus* (X). If a given GF network has an associated expression, it is marked by Y (else it is marked by N). Relevance Score (RS) represents the overall PMID weight associated with an article from where the protein was derived and the last column shows the number of articles in which the protein was present. As can be seen from Table 8, two proteins from the proteomics data were also extracted from the literature.

Five GF subnetworks were identified for axolotl: HDGF, TYMP, VEGFA, PDGFA, and PDGFB (all of these were literature-derived proteins). Ten GF subnetworks were identified for *Xenopus*: FGF2, GRN, HDGF, NGF, NOV, PDGFA, PDGFB, TYMP, VEGFA, and IGF1. *Xenopus* was found to contain all the growth factors present in the axolotl data. Two of these (GRN and PDGFB) were identified in the proteomics data.

GF Subnetwork	Axo(A)/ Xeno(X)	Expression	RS	No. of articles	p-Value (Axo)	p-Value (Xeno)
FGF2	X	N	13.76	1128		0.004341
GRN	X	Y (Xeno)	7.37	11		1.95E-76
HDGF	A/X	N	5.28	5	2.54E-24	8.54E-69
NGF	X	N	10.63	882		0.004341
NOV	X	N	8.69	13		5.38E-04
PDGFA	A/X	N	6.79	35	8.48E-04	5.38E-04
PDGFB	A/X	Y (Xeno)	6.72	17	8.48E-04	2.46E-04
TYMP	A/X	N	8.42	16	0.013574	6.05E-07
VEGFA	A/X	N	9.80	1410	0.011297	0.003863
IGF1	X	N	9.80	413		0.019281

Table 8. Growth factor subnetwork summary

FGF2 and VEGFA were present in 1128 articles and 1410 limb-regeneration specific articles respectively were the most commonly associated GFs with limb regeneration in the published literature. Both of these proteins are known to play a crucial role in limb regeneration. The p-values were calculated to determine the significance of subnetworks. In axolotl, HDGF was the most significant subnetwork (determined mainly by the number of genes from proteomics data which were present on a subnetwork, see Methodology section for calculation details).

While the role of HDGF or Hepatoma-derived growth factor is not very well-known in limb regeneration, it is highly expressed in tumor cells and known to play an important role in cancer progression. It is also very closely related to hepatocyte growth factor (HGF). In gastric carcinoma it has been shown that HGF regulates HDGF which in turn induces VEGF [251]. HDGF upregulation is also known to play an important role in liver regeneration [252]. Recently, in our segment defect regeneration study HGF was found to be of high importance and we showed improved bone regeneration across a critical size defect with HGF (results discussed in a later section).

Growth factor subnetworks were enriched for biological processes and pathways. Table 9 and 10 show the top 10 most enriched BPs and pathways respectively for the HDGF growth factor subnetwork in axolotl. The BPs and pathways represented below do not just represent HDGF but all the proteins in the subnetwork. The biological process category was derived from Gene Ontology (level 3) and pathways were derived from KEGG pathways.

Biological process	Number of overlapping proteins	p-Value
Regulation of cellular process	65	6.6E-220
Signal transduction	30	9.4E-145
Cellular macromolecule metabolic process	75	6.6E-136
Protein metabolic process	72	2.97E-87
Cellular nitrogen compound metabolic process	53	4.18E-84
Cellular biosynthetic process	50	7.06E-83
Regulation of metabolic process	53	8.42E-81
Transport	19	7.75E-79
Regulation of cellular metabolic process	51	1.95E-76
Nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	53	9.48E-74

Table 9. Biological process enrichment for HDGF subnetwork in axolotl

Pathways	p-Value
Olfactory transduction	4.38E-41
Cytokine-cytokine receptor interaction	1.17E-27
Neuroactive ligand-receptor interaction	1.67E-24
Regulation of actin cytoskeleton	8.35E-22
MAPK signaling pathway	7.68E-21
Pathways in cancer	4.51E-20
Chemokine signaling pathway	6.39E-16
Endocytosis	1.96E-15
Alzheimer's disease	2.86E-15
Focal adhesion	8.96E-15

Table 10. Pathway enrichment for HDGF subnetwork in axolotl

Seventy TF subnetworks were identified for the axolotl and a total of 158 such subnetworks were identified for *Xenopus*. This showed a very high representation of TFs in the limb regeneration data. Overall, 58 TFs were common in both the datasets. Of the 70 TFs identified from the axolotl network, nine were present in the proteomics data and the rest were derived from literature. The nine TFs from proteomics data were: ATF1, E4F1, HR, NFATC4, SOX6, TAF4, AHCTF1, ZNF592, and NEUROD2. Twelve TFs were uniquely identified in the axolotl data (AHCTF1, TAF4, ATF1, ZNF592, E4F1, HR, TLX1, TAF8, KLF10, NEUROD2, INSM1, and TAL1) and 100 TFs were unique to the *Xenopus* data. Of the 12 TFs unique to axolotl data, seven were from the proteomics data. Of the 158 TF subnetworks identified in *Xenopus*, 13 TFs were identified from the proteomics data.

Among all the TF subnetworks, 62 and 148 subnetworks contained at least two proteins from the proteomics data in axolotl and *Xenopus* respectively. Fifty-six and 100 subnetworks had p-value less than 0.05 in axolotl and *Xenopus* respectively. Table 11 (axolotl) and Table 12 (*Xenopus*) summarize the twenty most significant TF subnetworks. The number (No.) of proteins from the proteomics experiment and size of the subnetwork determine the significance or p-value. The p-value of the top two subnetworks in the *Xenopus* table is zero since among all the proteins on the subnetwork most of the proteins are from expression data or from proteomics experiments. This also indicates the high importance of such subnetworks since they contain many proteins from the proteomics proteins.

Among the important TFs identified, Myc is involved in various biological processes such as proliferation, growth, apoptosis, energy metabolism and differentiation [253, 254]. It has been shown to act with β-catenin to inhibit wound healing by interfering with differentiation in chronic ulcers [255] and is expressed in regenerating limb and lens of the newt. In the newt Notophthalmus *viridescens*, in-situ hybridization has shown that Myc is localized in both the epidermis and subjacent blastema cells. This expression has been correlated with the maintenance of blastema cell proliferation [256, 257]. Recently, along with other stem cell factors, Myc expression in *Notophthalmus viridscens* was found to be highest during the dedifferentiation phase of blastema formation. Expression then decreased at later stages but still remained higher than the control tissue [258]. These studies have related Myc to proliferation as well as stemness but the downstream targets of Myc which result in these effects have not been identified. The Myc subnetwork identified here is connected to 38 proteins from the axolotl proteomics data and 176 proteins from the proteomics data. Our previous analysis had also shown the importance of Myc in the axolotl data [6]. Based on this information, the specific roles of Myc in limb regeneration should be explored further.

TF Subnetwork	No. Of Proteins On Subnetwork	No. Of Proteomics Proteins	p-Value
AHCTF1	55	45	1.14E-71
TAF4	50	32	4.64E-45
ESR1	235	50	7.72E-41
ATF1	32	23	1.82E-34
SOX6	33	23	5.90E-34
ZNF592	29	21	1.16E-31
TP53	268	44	7.77E-31
MYC	207	38	1.52E-28
SP1	72	21	1.05E-20
YBX1	91	22	1.13E-19
ATF2	105	23	1.81E-19
E4F1	22	12	2.48E-16
TP63	54	16	3.49E-16
POU5F1	66	17	5.59E-16
NOLC1	81	17	2.24E-14
ERG	67	13	7.69E-11
RELA	25	9	1.62E-10
AIRE	52	11	8.52E-10
TCF3	35	9	4.73E-09
SMARCA4	42	9	2.66E-08

Table 11. Significant TF subnetworks in axolotl

Specificity Factor1 or SP1 is a ubiquitously expressed protein and has varied roles in cell growth, differentiation, apoptosis, angiogenesis, tumorigenesis and immune response. It is known to interact with cyclins which promote the G1/S phase transition, as well as with cyclin-dependent inhibitors that inhibit progression through the cell cycle. Similarly, its target genes include both pro- and anti-apoptotic genes and pro- and anti-angiogenic genes. Specificity factor1 is also linked to chromatin remodeling through its interaction with p300 and histone deacetylases. Specificity factor1 is known to interact with several TFs including Myc in order to activate several downstream targets. However, SP1 action is highly dependent on its interaction with other members of the SP family and extracellular signals [259-261]. In this analysis, SP1 connects to 21 proteins from the axolotl proteomics data and 40 proteins from the *Xenopus* proteomics data (data not shown in the tables here). Our previous network analysis had also found strong evidence for the involvement of SP1 in limb regeneration [6].

TF Subnetwork	No. Of Proteins On Subnetwork	No. Of Proteomics Proteins	p-Value
E2F1	346	278	0
ILF2	592	336	0
NCOR1	374	252	2.47E-282
TFAP2C	198	188	1.47E-263
BTF3	160	143	7.10E-188
SALL1	138	133	1.81E-186
BRD7	149	128	2.01E-162
TP53	750	205	2.55E-123
MYC	541	176	9.50E-119
ESR1	541	170	1.56E-111
BANP	78	65	8.66E-80
NOLC1	262	100	1.90E-73
YBX1	313	103	3.74E-68
ETV3	54	50	1.04E-66
ATF2	276	93	1.83E-62
TP63	129	67	3.64E-60
POU5F1	120	65	5.65E-60
ERG	178	70	2.73E-52
TCF3	111	50	4.72E-41
NFIA	77	42	3.72E-39

Table 12. Significant TF subnetworks in *Xenopus*

Other TFs such as msx-1, nrad, Klf4, Oct4, Sox2, and Lin28 are associated with stemness and are expressed during formation of the accumulation blastema [7, 258, 262-266]. Among these, this analysis identified Oct4 (POU5F1) Sox2 subnetworks from both axolotl and *Xenopus* networks. Recently, combinations of the TFs Myc, Oct4, Sox2, Klf4, Lin28, and Nanog were shown to reprogram adult fibroblasts to iPSCs [267, 268]. c-Myc has been shown to enhance the ability of Oct4, Sox2 and Klf4 to induce pluripotency up to 10-fold [267]. However, high levels of Myc are only transiently required and sustained levels were found to lead to tumors [253, 254]. C-myc, Klf4 and Sox2 have been shown to be expressed in regenerating newt limb tissue and Lin28 in regenerating axolotl limb tissue [7, 256-258]. Figure 12 in the previous section shows the network linking these important TFs. These findings suggests that the TFs identified in these subnetworks (especially Myc, SP1, Oct-4, Sox2) are central to a network of TFs that regulate mesenchymal stem cell properties of the blastema and that play a role in the nuclear reprogramming of differentiated limb cells to blastema cells.

The subnetworks were also extracted for the ECM protein molecular class. Twenty-two ECM protein subnetworks were identified for the axolotl and 31 for *Xenopus*. Interestingly, 11 of the 22 axolotl ECM proteins were identified from the proteomics data and 14 among 31 ECM proteins in *Xenopus* data. This suggested a very high enrichment of ECM category in the proteomics data. Among all three molecular classes analyzed for limb regeneration, ECM had the highest representation from the proteomics data. Seven ECM proteins were unique to the axolotl data while 16 were unique to the *Xenopus*. Table 13 and Table 14 show the most significant ECM subnetworks in axolotl and *Xenopus* respectively.

ECM Subnetwork	No. Of Proteins On Subnetwork	No. Of Proteomics Proteins	p-Value
FN1	605	149	1.66E-143
COL1A1	45	26	5.01E-35
LTBP4	16	10	1.23E-14
VTN	24	7	9.85E-08
MATN2	8	4	5.33E-06
DCN	13	4	5.01E-05
COL6A1	3	2	8.48E-04
ELN	3	2	8.48E-04
MATN1	3	2	8.48E-04
MATN4	3	2	8.48E-04
NID1	3	2	8.48E-04
THBS1	3	2	8.48E-04
AGRN	4	2	0.001668
EFEMP1	4	2	0.001668
HSPG2	4	2	0.001668
LAMB1	4	2	0.001668
COL7A1	5	2	0.002733

Table 13. Significant ECM protein subnetwork statistics for axolotl

Among the subnetworks, FN1 and COL1A1 were the most significant subnetworks in axolotl. FN1 was also the most significant subnetwork in *Xenopus*. Fibronectin 1 (FN1) and COL1A1 were both found to be expressed in the proteomics data as well. FN1 was also found to be expressed in the *Xenopus* proteomics data. Among all the three molecular classes analyzed, FN1 was found to have the highest coverage or highest number of connections with the proteins from the proteomics data. In axolotl, a total of 263 proteins were present on the network and of these 149 were connected to FN1 (at most two steps away from FN1 on the subnetwork). In *Xenopus*, 601 proteins from the proteomics data were present on the network and 402 of these were present in the FN1 subnetwork.

ECM Subnetwork	No. Of Proteins On Subnetwork	No. Of Proteomics Proteins	p-Value
FN1	1150	402	0
MFAP1	216	190	8.76E-250
THBS1	82	63	5.13E-73
COL1A2	19	14	1.47E-16
VTN	27	13	4.69E-12
COL1A1	10	6	6.01E-07
COL2A1	6	4	3.12E-05
THBS2	7	4	7.00E-05
AGRN	4	3	2.24E-04
HSPG2	4	3	2.24E-04
COL3A1	10	4	3.73E-04
COL5A1	5	3	5.38E-04
CYR61	6	3	0.001035
MATN2	7	3	0.001742
TFIP11	17	4	0.003212
COL6A1	3	2	0.004341
DCN	3	2	0.004341
ELN	3	2	0.004341
NID1	3	2	0.004341

Table 14. Significant ECM protein subnetwork statistics for *Xenopus*

In the axolotl proteomics data, components of collagen 1 were upregulated at all or two of the three time points. Components of cartilage matrix (collagen 2) and basement membrane (collagen 4) were downregulated at all dpa, as was decorin, which interacts with collagen1 fibrils and may affect the rate of their formation. However, MATN 4, a major component of cartilage matrix, was upregulated at 1 and 4dpa, then downregulated at 7dpa. FBN1, a large glycoprotein that associates with elastin to provide force-bearing support in the ECM, was upregulated at 1 and 7dpa, with no change at 4dpa. MATN 2, a von Willebrand family member involved in matrix assembly, was upregulated at 1 and 4dpa, then returned to control level at 7dpa. FN1 forms part of the provisional wound matrix (clot) and was upregulated at all dpa.

The upregulation of FN1 and collagen 1, the downregulation of collagens 2 and 4, and the downregulation of EHD4, an endosomal trafficking regulatory protein [269] present in the matrix of differentiating cartilage and fibroblastic connective tissue during rat limb development [270], is consistent with other observations indicating that the differentiated tissue matrix is replaced by an ECM that is more similar to the limb bud matrix, and more favorable to the migration of dedifferentiated cells to form the blastema under the wound epidermis [271]. The proteomics data obtained for FN1 in axolotl was validated by immunostaining method in our previously published study [7]. Based on the subnetwork analysis and the analysis reported in the proteomics data, FN1 is a promising target for limb regeneration and its role should be further explored.

Differential Subnetwork Analysis

The approach developed here (as outlined by the Algorithm 3.6 in the Methodology section) compared all the subnetworks across two given conditions instead of comparing individual genes. The differential subnetworks were evaluated on both topological and biological properties associated with the nodes of the subnetworks. A significant outcome of this approach was the ability to identify differences between the same protein in two conditions based on its interaction profile, expression pattern, function and pathway differences. This approach was designed to include the direct neighborhood of the node—not just the node itself—to compare all the features (other than expression pattern) for subnetwork comparison. Biologically, this is very important since it signifies differential connectivity of the same protein between two conditions. It should also be noted that a mathematical comparison of condition 1 with condition 2 is different from condition 2 with condition 1. It is important that the research question is clearly described before differential comparison. If the interest is to find differences in condition 1, then the first case described above is used while if the interest is to find differences in condition 2, then the second case is used. In other words, either condition 1 or 2 are used as the base for evaluating condition-specific differences.

The differential subnetwork algorithm 3.6 described in the Methodology section generates an excel file, Disco.xls (**Di**ssimilarity **Sco**re - DS) which contains the DS for all the common nodes between the subnetworks being compared. Each subnetwork from one condition is compared with all those subnetworks in the second condition which have one or more common nodes. The scores of these individual node comparisons are averaged to generate an overall DS for the subnetworks. The range of the DS is from 0 to 1 with 1 being the highest dissimilarity and zero being no evidence for dissimilarity. However, it should be noted that once the differential subnetworks of interest are identified, the DS at the node level can be used to identify the most differential nodes or proteins which can be used as targets (or key proteins) to influence the subnetwork of choice. In other words, these key proteins can be used to design further biological experiments (such as knock-down). Hence, this strategy not only provides an informed decision about the most differential components between the two biological systems, it also helps identify the targets which can then be used to stimulate a desired biological response.

The limb regeneration dataset was compared for all three molecular classes for which subnetworks were identified: GFs, TFs, and ECM proteins. The goal of this analysis was to identify differential subnetworks that are instrumental in conferring regeneration ability in the regeneration-competent axolotl. So, the comparisons were performed such that differences in axolotl were highlighted as opposed to *Xenopus* (first case according to the description above, axolotl subnetworks refer to condition 1 and *Xenopus* subnetworks refer to condition 2).

Growth Factor Subnetwork Comparison

Each of the five growth factor subnetworks identified for axolotl (described above) were compared with each of the ten GF subnetworks for *Xenopus*. HDGF was the most significant GF subnetwork containing 108 total proteins in axolotl and 228 proteins in *Xenopus*. Comparison of HDGF subnetwork in axolotl and *Xenopus* showed that seventy seven proteins were common between these subnetworks. Among these 77 common proteins, 27 proteins were from the proteomics data, 75 had known gene ontology biological processed and 48 proteins were involved in known pathways. An overall dissimilarity score for the comparison of HDGF subnetworks was 0.33 indicating more similarity between the two subnetworks than dissimilarity. However, eight proteins in the HDGF subnetwork comparison had a very high DS as reported in Table 15. Column 2-6 in Table 15 reflect the dissimilarity score for each of the factors that were evaluated in the differential subnetwork algorithm. N/A: missing values for that factor.

The proteins derived from literature (missing expression values) are highlighted in yellow.

Gene Symbol	Interaction Dissimilarity	Expression Dissimilarity	GO Dissimilarity	KEGG Dissimilarity	Total Dissimilarity
HNRNPU	0.8	1	0.77	0.95	0.88
FBL	0.76	1	0.7	0.75	0.8
SND1	0.77	1	0.66	0.75	0.79
TRIM24	0.67	N/A	0.71	1	0.79
RPL4	0.63	1	0.51	0.8	0.73
SMARCA4	0.8	N/A	0.31	1	0.7
IL7R	0.73	1	0.32	0.5	0.64
SIRT7	0.44	1	0	1	0.61

Table 15. Proteins with high dissimilarity scores in HDGF subnetwork comparison

Among the eight proteins with a high DS, two proteins, TRIM24 and SMARCA4 were derived from literature (also note that HDGF itself was derived from literature). Heterogeneous Nuclear Ribonucleoprotein U (HNRNPU) and Fibrillarin (FBL) were the most dissimilar proteins in the subnetwork comparison and both these proteins are involved in mRNA processing. TRIM 24, SND1 are related to transcriptional control and SMARCA4 is involved in chromatin modification which is required for transcriptional activation. This siginifies that mRNA processing is a critical level of control for protein synthesis in general during limb regeneration. It is also evident that the most significant differential components of the HDGF subnetwork are the proteins with a high dissimilarity score.

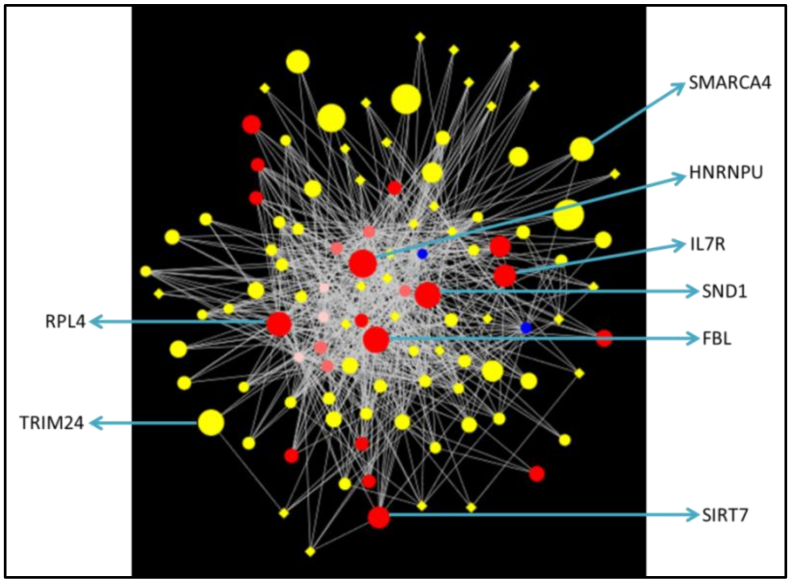

Figure 17. HDGF subnetwork of axolotl highlighting dissimilarity

Figure 17 shows the HDGF subnetwork and significant proteins discussed in Table 15 are labeled in the figure. Figure 17 also shows the variability in the DS of the different nodes on the subnetwork. Please note that diamond shaped nodes (unique to the axolotl subnetwork are small since DS was evaluated for the nodes common to both the axolotl and *Xenopus* subnetworks).Node size reflects the dissimilarity score (the bigger the node, the higher the dissimilarity score), red and blue node color are indicative of dissimilarity in expression values (bright red is highly dissimilar – dissimilar pattern at all time point comparisons, blue color – nodes had a similar pattern of expression), yellow colored nodes were derived from literature and node shape indicates common (round) or unique nodes (diamond) on this subnetwork.

Transcription Factor Subnetwork Comparison

Seventy transcription factor subnetworks of axolotl were compared with 158 TF subnetworks of *Xenopus*. Table 16 below shows the overall DS of the highly significant TF subnetworks as identified above. Most of these subnetworks are also known to influence stemness and hence are crucial to limb regeneration (Myc, POU5F1, SP1, SOX2). TF subnetworks were discussed above. Most of these subnetworks were found to be huge (high number of nodes) and yet around 50% dissimilar between the axolotl and *Xenopus*, highlighting the potential differences in the subnetworks between these conditions. SOX2 was identified as the most dissimilar subnetwork among the subnetworks mentioned in Table 16.

TF Subnetwork	No. of common proteins	Total Proteins	DS
MYC	190	207	0.5
ESR1	210	235	0.53
SOX2	19	22	0.58
POU5F1	65	66	0.42
SP1	59	72	0.57
TP53	248	268	0.48

Table 16. Overall dissimilarity score of the highly significant subnetworks

TP53 was the subnetwork with the highest number of total nodes (268) followed by ESR1 (235) and Myc (207). It is evident from Table 16 that subnetwork size is not indicative of the DS. All these subnetworks have a very high number of common nodes and are still dissimilar! These results also provide a validation to our initial hypothesis that the nodes between two conditions might be similar but it is the differential connectivity between these nodes which is crucial for the given biological condition.

To identify the most important differentially connected nodes, Figure 18 and Figure 19 were constructed to show the subnetwork comparison between the same TF subnetworks in axolotl and *Xenopus*. The common nodes or proteins between the subnetworks in axolotl and *Xenopus* are represented on the y-axis. As an example, the first column in Figure 18 represents the common proteins for the Myc subnetwork in axolotl and *Xenopus*. The colors indicate the dissimilarity score (bright yellow color indicates zero dissimilarity or similarity and bright red color indicates high dissimilarity) for the same node comparison in axolotl and *Xenopus*. The black color indicates that those nodes were not present in that subnetwork. The names of the most dissimilar nodes are labeled on the y-axis. Figure 18 shows the hierarchical clustering for the nodes or proteins on Myc, ESR1 and TP53 subnetworks.

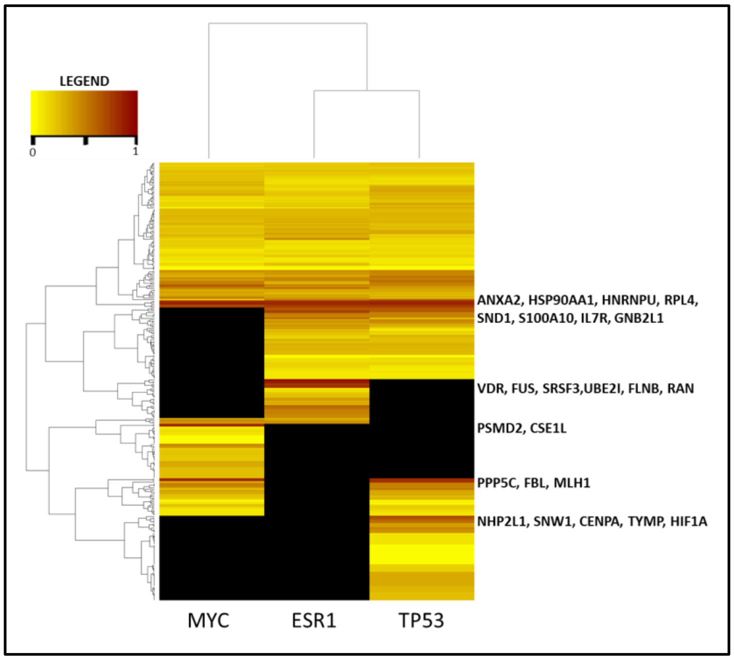

Figure 18. Hierarchical clustering of the subnetworks with respect to dissimilarity scores

ESR1 and TP53 were the most closely related subnetworks (clustering on the x-axis) — this is also because of the high number of shared nodes with a similar DS between these subnetworks. It can also be seen in the Figure 18 that although there are some nodes which are very similar (yellow color tends to be closer to DS of zero, as indicated by the legend image in the figure) between the two conditions, there are a lot of nodes with high dissimilarity (red color). The nodes with the highest DS are indicated on the y-axis and these are the most significant differential components of these subnetworks. These nodes also can be considered as key proteins responsible for conferring limb regeneration ability in axolotl.

Some of these nodes such as HNRNPU, SND1, are similar to the ones identified as important for the HDGF subnetwork comparison—hence, indicating a potential role of these proteins in controlling different processes within these systems. Of particular interest are ANXA2 and S100A10 which are known to be involved in limb regeneration by controlling the immune response [218, 221, 272, 273]. Higher levels of several S100 family Ca^{2+}-binding proteins are observed in the regenerating ear tissue of MRL/MpJ-Fas mice vs. non-regenerating ear tissue of C57BL/6J mice, as determined by laser capture proteomics [274, 275]. In axolotl proteomics data, ANXA2 was found to be upregulated at 1 and 4dpa. ANXA2 is an autocrine factor that promotes osteoclast formation and bone resorption.

Figure 19 shows the hierarchical clustering of nodes in the SOX2, SP1, and POU5F1 subnetworks between the axolotl and *Xenopus* data. It should be noted that similar to Figure 18, these are the results for the same subnetwork comparison in both conditions (as an example, SOX2 subnetwork in axolotl when compared with SOX2 subnetwork in *Xenopus*).

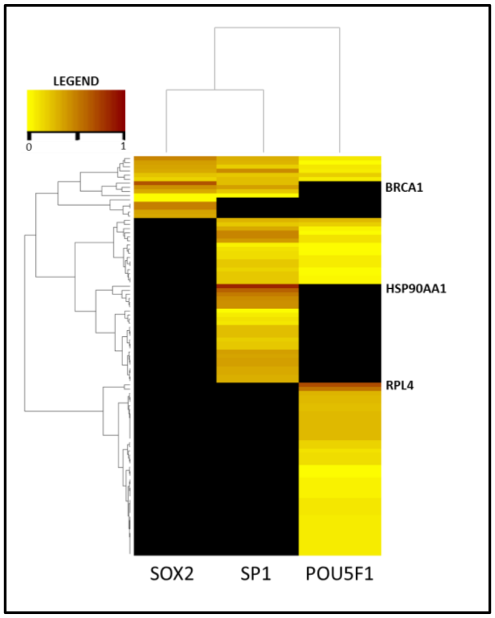

Figure 19. Hierarchical clustering of the subnetworks with respect to dissimilarity scores

SOX2 contained the least number of total nodes and POU5F1 subnetwork contains nodes with much higher similarity as compared to the other subnetworks. The nodes BRCA1 (highly dissimilar in SOX2), HSP90AA1 (highly dissimilar in SP1), and RPL4 (highly dissimilar in POU5F1) were the most important key molecules in these subnetworks. Of these, HSP90AA1 was also detected in the axolotl and *Xenopus* proteomics data. It was downregulated in the axolotl data while upregulated in the *Xenopus* data.

Table 17 below shows the most dissimilar TF subnetworks in axolotl as compared to *Xenopus*. Overall, NFATC4 and SAFB were the most dissimilar subnetworks. To obtain these results, the subnetworks with less than five proteins were not included in the analysis.

TF Subnetwork	No. of common proteins	Total Proteins	DS
NFATC4	7	8	0.8
SAFB	12	15	0.74
NFIA	11	12	0.66
HIF1A	21	23	0.65
HR	8	14	0.65
EPAS1	13	13	0.64
SSRP1	28	30	0.64
HSF1	11	15	0.63
TCF4	8	8	0.63
BCL6	8	10	0.61
CEBPA	6	9	0.6
STAT3	7	7	0.6
ZNF592	19	29	0.6

Table 17. Transcription factor subnetworks with high dissimilarity

Extracellular Matrix Protein Subnetwork Comparison

The extracellular matrix protein subnetworks were investigated similar to the GF and TF subnetworks described above. Table 18 shows the overall dissimilarity for the ECM subnetworks. None of the ECM subnetworks with less than a total of 5 proteins were included in the table. Among the ECM subnetworks, MATN2 was the subnetwork with the highest overall DS. However, it should be noted that the number of total nodes in this subnetwork is relatively very low as compared to the other subnetworks and hence indicating a low connectivity with other subnetworks.

ECM Subnetwork	No. of common proteins	Total Proteins	DS
FN1	507	605	0.4
COL1A1	36	45	0.49
VTN	23	24	0.48
LTBP4	9	16	0.49
DCN	12	13	0.61
COL2A1	9	11	0.68
MATN2	7	8	0.71

Table 18. Extracellular matrix protein subnetworks with high dissimilarity

Gene Symbol	Interaction Dissimilarity	Expression Dissimilarity	GO Dissimilarity	KEGG Dissimilarity	Total Dissimilarity
PCMT1	0.86	1	0.97	1	0.96
TPM3	0.88	1	0.96	1	0.96
ACTN4	0.86	1	0.9	1	0.94
ACACA	0.84	1	0.74	1	0.9
FUS	0.82	1	0.79	1	0.9
KHSRP	0.84	1	0.74	1	0.9
HNRNPU	0.81	1	0.79	0.95	0.89
PELP1	0.75	N/A	1	N/A	0.88
HIST2H2BE	0.86	1	0.68	1	0.88
ANXA2	0.8	1	0.71	0.95	0.86
SRSF3	0.84	1	0.58	1	0.86
PSMD2	0.8	1	0.73	0.93	0.86
C1QA	0.67	N/A	1	N/A	0.84
SEPT9	0.67	N/A	1	N/A	0.84
EPPK1	0.67	1	0.7	1	0.84
PTBP2	0.67	N/A	1	N/A	0.84
FLNB	0.76	1	0.63	0.92	0.83
GSTP1	0.81	1	0.83	0.67	0.83
HSP90AA1	0.86	1	0.55	0.92	0.83
ETF1	0.75	1	0.75	N/A	0.83
MME	0.64	1	0.67	1	0.83
CLTC	0.92	N/A	0.54	1	0.82
FBL	0.76	1	0.73	0.8	0.82
SND1	0.78	1	0.7	0.8	0.82
CACNA1A	0.67	1	0.79	N/A	0.82
USP39	0.75	N/A	0.67	1	0.81
FLOT2	0.6	N/A	1	N/A	0.8
NHP2L1	0.7	1	0.66	0.83	0.8
RPL4	0.65	1	0.67	0.88	0.8

Table 19. Proteins with high dissimilarity score (DS) in the FN subnetwork comparison

Interestingly, as mentioned earlier, the FN subnetwork was the most connected subnetwork with the highest number of nodes among all the subnetworks analyzed in this study. So, we compared the FN subnetworks between axolotl and *Xenopus* to evaluate its dissimilarity. Although its overall dissimilarity score is 0.4, there were a lot of very highly dissimilar nodes in this subnetwork (mentioned in the Table 19 above). These nodes with a high dissimilarity score should be further investigated for their role in limb regeneration.

Of particular interest was HNRNPU as it was identified as one of the most dissimilar nodes with multiple subnetwork comparisons between axolotl and *Xenopus*. It was also identified among the subnetworks of all the categories (GF, TF, and ECM) or molecular class of proteins that were analyzed. Figure 20 below shows the HNRNPU connectivity in both axolotl and *Xenopus* networks to highlight the presence of differential neighborhoods. It should be noted that the coloring of the nodes is similar to that described for condition-specific networks, pink and blue nodes were derived from literature while yellow and red nodes were derived from proteomics data. Nodes with the green boundary are the common nodes between axolotl and *Xenopus* data. Two important observations from this data are: (i) Although the overall axolotl network is significantly smaller than *Xenopus* network (refer Figure 16), HNRNPU neighborhood is far denser in axolotl as compared to *Xenopus*; (ii) Only 10 nodes are common between axolotl and *Xenopus*.

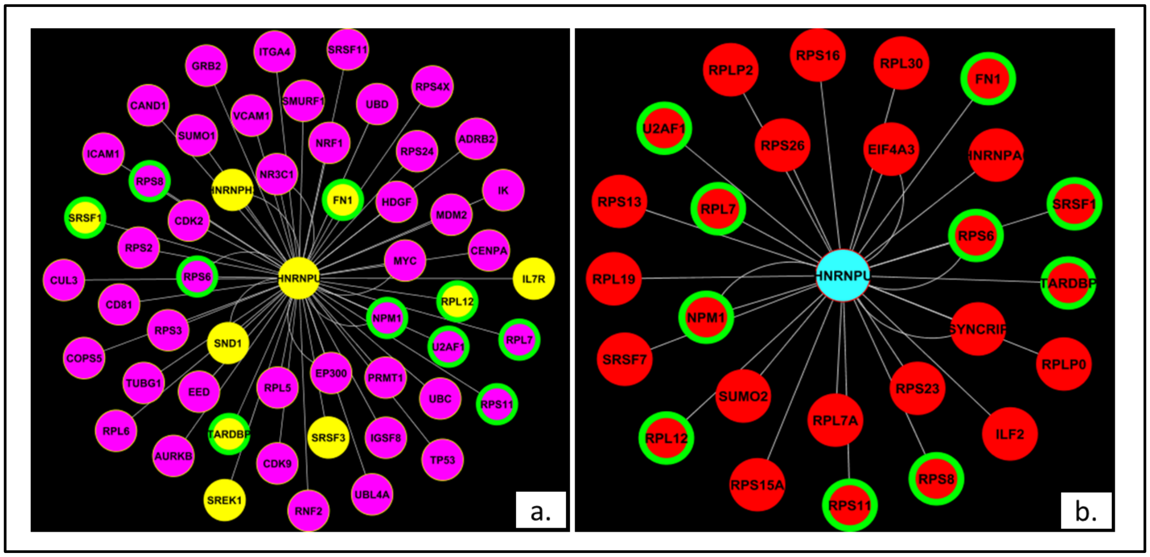

Figure 20. HNRNPU neighborhood in axolotl (a) and *Xenopus* (b) networks

These results signify the differential connectivity of common proteins such as HNRNPU. Hence, the role of these proteins in limb regeneration should be further evaluated. Importantly, this validates our hypothesis that although the proteins can be similar between the two biological conditions, their interacting partners could be different leading to physiological differences in between the two conditions. Such differential components can serve as important regulators governing a biological system and hence their role as potential targets should be further investigated.

Biological Validation

Proteomics Validation

We selected NOS1, fibronectin and α-actinin for validation of axolotl proteomics data by immunocytochemistry at 1 and 7dpa. The Figure 21 shows longitudinal sections of control (a,d,g) vs. 1dpa (b, e, h) and 7dpa (c, f, i) axolotl hindlimbs stained with primary antibodies to NOS1 (a-c), FN1 (fibronectin 1) (d-f), ACTN (α-actinin) (g-i).

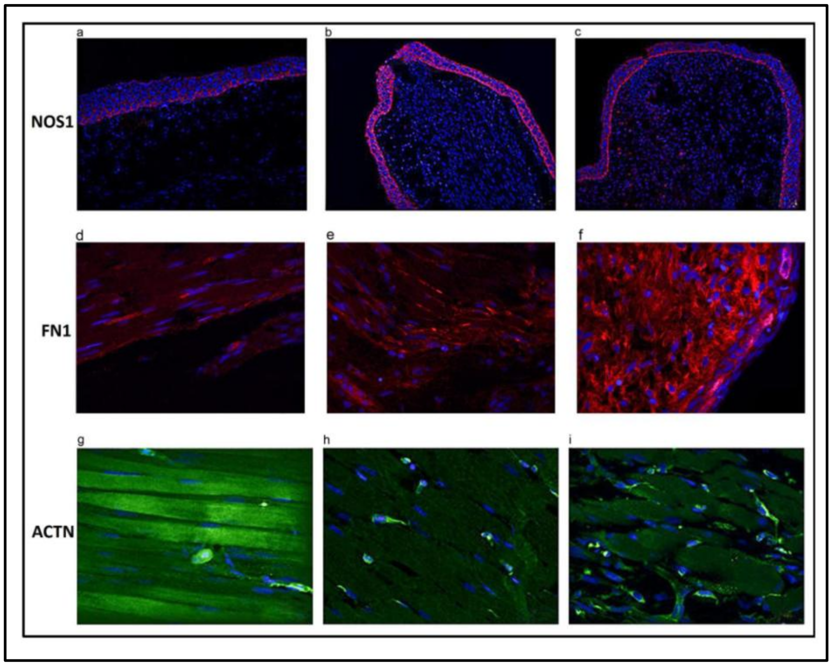

Figure 21. Immunostained sections of axolotl hind limbs

Conjugated secondary antibodies were alexa-568 for fibronectin and NOS1, and alexa-488 for α-actinin. Nuclei were counterstained with DAPI. As expected, fluorescence intensity of NOS1 and fibronectin staining (red) at 1 and 7dpa showed significant increases compared to controls, while α-actinin staining intensity (green) showed a significant decrease. The fold changes determined by LC/MS/MS were largely congruent with densitometric measurements, indicating that quantitative LC/MS/MS data accurately reflected the levels of specific proteins. We similarly validated the proteins from proteomics data in *Xenopus*. More details can be found in our previous publications [7, 226].

Segment Defect Regeneration

Fifty percent defects were chosen in order to provide a regenerative challenge well beyond the critical size defect (CSD). We noted that in many cases, the bone at the cut ends of the fibula had undergone substantial regression, making the segment defect closer to 70%.

The extent of regeneration fell into two categories, partial and significant. Partial regeneration was defined as bridging less than 25% of the defect, whereas significant regeneration was defined as bridging 50% or more of the defect. The 7-factor combination yielded one case of partial regeneration out of 24 limbs, and the BMP-4/VEGF combination yielded two cases out of 24 limbs. The regenerated skeletal tissue consisted of irregular tongues of cartilage. No cases of significant regeneration resulted from these combinations.

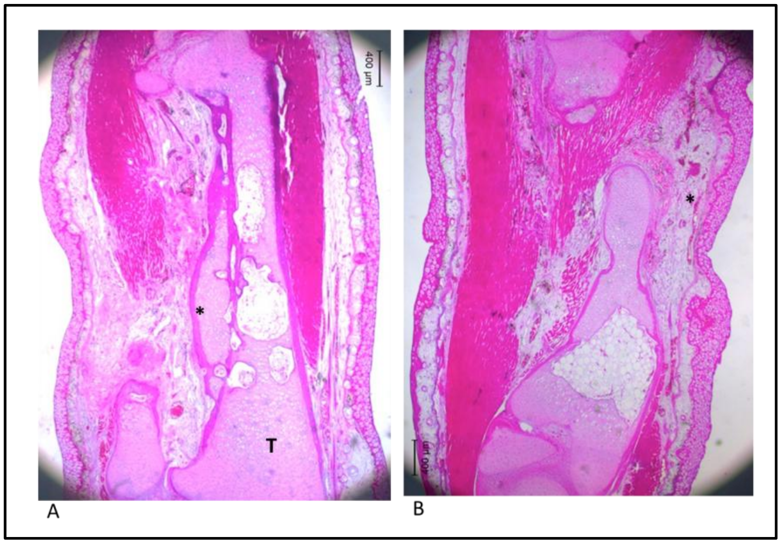

Figure 22. Two 50% defects treated with BMP4/HGF, three months post-operation

The BMP-4/HGF combination yielded two cases of partial regeneration and four cases of significant regeneration out of 24 limbs. BMP4/HGF induced significant regeneration in four out of 24 limbs by two-three months post-implantation. Figure 22A illustrates one case where new cartilage surrounded by a shell of bone regenerated the length of the gap, but in parallel with, and adjacent to, the tibia. The origin of the cells of this cartilage was likely the periosteum of the tibia. Figure 22B shows a case where regeneration took place over nearly the whole defect from the proximal end of the fibula. Figure 22A shows that an irregular secondary length of cartilage (asterisk) was induced along the axis of the tibia (T). Vertical lines indicate the boundaries of the gap in the fibula. No skeletal tissue was regenerated within the defect space itself except for a nodule of cartilage. Figure 22B shows cartilage (asterisk) from the distal end of the fibula regenerated across 80% of the defect.

The only combination that promoted significant regeneration of cartilage and bone across 50% segment defects with any consistency was BMP-4/HGF. Partial regeneration was stimulated in one case of the 7-factor cocktail and two cases of the BMP-4/VEGF combination, but none of the other combinations stimulated regeneration.

There could be multiple explanations for these differential results, such as suboptimal GF concentrations and/or concentration ratios. A likely part of the explanation, however, is that as in fracture repair, the expression of different GFs needs to follow a spatial and temporal cascade initiated by BMPs in order to regenerate across a CSD. The involvement of HGF would be through its ability to induce expression of BMP receptors. The release kinetics profile shows that after a peak burst at 2 hr and a subsequent 15% decrease by 4 hr, the amount of BMP-4 released is sustained at a relatively steady level of about 75% of the 2 hr value over three days. We did not test BMP-4 or HGF alone, so the possibility remains that either of these GFs could by themselves initiate the molecular cascade leading to cartilage regeneration.

In BMP-4/HGF-treated 50% defects, as in untreated 10% and 20% defects, the regenerating cartilage appears to grow from either or both cut ends of the fibula, suggesting either a periosteal or chondrogenic origin. More details can be found in our work which will be published soon [245]. We have established that the axolotl (and most likely other urodele species) can serve as an inexpensive and surgically amenable model to screen different combinations of factors for their ability to promote regeneration of cartilage and bone across a CSD. The model has established that a combination of BMP-4 and HGF, as well as whole limb issue extract is effective in evoking regeneration across gaps of 50% or greater in the axolotl fibula.

CHAPTER FIVE: CONCLUSIONS

Summary

To address the challenges in the identification of condition-specific differential components of a biological system, we developed a novel and innovative systems level approach to identify the differential subnetworks and key target molecules. This approach provided a strategy to not only prioritize and discover differential components from high throughput experiments but also identified condition-specific data from the published literature.

Condition-specific ontologies along with a probabilistic model for prioritization of relevant articles were used to mine the published literature. The literature-derived data was then combined with the experimentally derived proteomics data to construct condition-specific protein interaction networks. These networks were then used to derive the molecular class based subnetworks for each condition. These subnetworks were further compared by incorporating both the biological and topological properties of the nodes (proteins) and edges (interactions) in the model to identify the differential subnetworks.

This approach was implemented to understand the differences between the limb-regeneration competent system of axolotl and. the limb-regeneration deficient system of *Xenopus*. Limb regeneration specific articles were mined from the published literature and assigned a relevance score. Proteins were then extracted from the articles with a significant relevance score. The proteins derived from the proteomics data collected at different time points after amputation were combined with the literature-derived proteins to construct competent and deficient networks. These networks were then used to identify growth factor, transcription factor, and extracellular matrix protein subnetworks. The subnetworks were then further compared to identify most dissimilar subnetworks and key proteins that possibly confer the ability to regrow the limbs in the competent system of axolotl. Key growth factors identified for segment defect regeneration were biologically validated by loading them onto scaffolds specifically designed to deliver these growth factors in the critical size defect models of axolotl. We observed an increased regenerative response with this approach as compared to the controls. The biological experiments validated our in-silico model approach to identify significant growth factors from literature-mined data and network analysis.

Limitations and future work

The approach designed and implemented in this study suffers from some limitations which can be improved in the future. One of the limitations of this approach is the use of experimental data for analysis (microarray/proteomics). Several different and more complicated datasets are being produced by the biological/bioinformatics community such as epigenetics, next generation sequencing (NGS) data, etc. At present, this methodology does not support integration of these several different datasets for a differential comparison. However, we believe that the additional datasets can be easily added as another feature in the model to estimate dissimilarity. To demonstrate the easy adaptability and working of the differential network analysis with the next generation sequencing data, we analyzed the whole genome, sequencing data of 37 Korean individuals and applied the differential subnetwork algorithm in order to understand the conserved modules among family members.

Most of the NGS studies have focused on the upstream analysis of the data. In this study, we provided a framework for analyzing the WGS data using systems biology approaches to identify the significant functional components. The pipeline was built on the single nucleotide variants from the Korean Personal Genome Project (KPGP) dataset and identified 1.4M low frequency variants and 1.3M novel variants. Function and pathways analysis, and significant modules in the KPGP variant gene network showed an enrichment of complex diseases like cancer and neurodegenerative disorders. This study also identified the highly conserved modules within the family members. Figure 23 shows the conserved modules identified in this study. Darker color indicates high similarity/low dissimilarity while bright yellow color indicates higher dissimilarity.

An overall trend emerges from Figure 23, dissimilarity in a family is the lowest for twins, followed by sibling or parent-child combinations, then cousins and lastly the unrelated members of a family (both the parents). The multicultural family comparison between the mother (KPGP10) and her children (KPGP11/12) was the most striking in terms of a very high DS. This could be due to the more dominant genotype of the father (KPGP9) with whom the children share a very high similarity (in accordance with other parent-child pairs of the same family). However, more in-depth analysis is required to understand such variability. The most significant modules among the twins were the CHEK2 and COPS7B which are involved in cancer [276] and neurodegeneration [277] respectively. The least dissimilar modules overall for all sample comparisons were CEP170, CHEK2, MLL3, and PDE4DIP modules. MLL3 possesses histone methylation activity and is involved in transcriptional coactivation [278]. MLL3 along with other genes in the module are known to be involved in several kinds of cancers, development disorders and brain-related malfunctions such as ALS. CEP170 is also involved in similar disorders [279]. Please refer to our publication for more details [280]

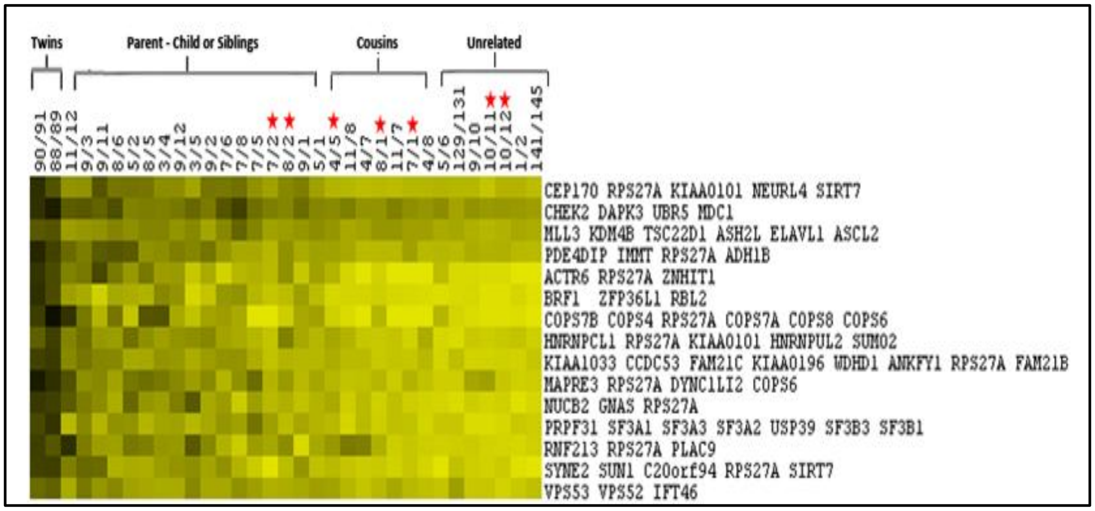

Figure 23. Heatmap for the family comparisons showing conserved modules

Another limitation of this approach is that the ontologies used to determine the CL can be queried through SPARQL to provide a more efficient method for querying. Several levels within the ontologies can then be specifically queried to resolve synonyms and get rid of more general science words (which were manually removed in the present approach). To improve the efficiency of protein extraction from relevant articles, a better methodology can be designed to retrieve full names by using partial matching. However, the present implementation of partial matching yields a lot of spurious results and so the present approach used only exact dictionary matching. Moreover, the protein interaction network construction relied on the data present in BioGRID [22]. BioGRID is one of the biggest repositories of protein-protein interactions (PPIs) [281]. It holds 696,237 interactions for 46 organisms which are derived from 40,858 publications. A team of 14 curators manually curates this vast amount of literature [22]. This and other manually curated databases also store information about protein interaction detection methods (PIDMs, such as yeast two hybrid, co-immunoprecipitation, etc.). A simple search on PubMed with the term “protein protein interactions” returns about 279,556 articles. So far, the number of publications manually curated by all the databases put together is not more than 60,000. There is a huge gap between the manually curated articles and those contained in PubMed. Such unannotated publications contain valuable information on PPIs and PIDMs that can be useful to the scientific community.

To reduce the knowledge gap between the number of articles in PubMed containing PPIs and PIDMs vs. the number of articles in publically available databases, it is essential to develop efficient data mining methods. Although BioCreative III tasks identified several ways to solve this issue, the performance of these methods was quite low as described above. This was mainly because the task was to identify most of the known interaction detection methods (almost 115). However, most of these detection methods are now obsolete and not considered to be significant. The three methods, co-immunoprecipitation (anti-bait and anti-tag), pull-down, and yeast two hybrid constitute almost half of the methods present in documents made available by BioCreative. These are also the methods considered to be most significant in the detection of protein interactions by the scientific community in general.

We developed a methodology for the identification of these three most significant methods for protein interaction detection. We argued that instead of treating the PPI and PIDM as two different tasks (as defined in the BioCreative challenge), both should be used together to identify significant PPIs. This is because many articles in the biomedical literature contain PPI sentences (usually referring to already known interactions) but that does not necessarily indicate that those PPIs were detected by some biological technique in that study. Such studies cannot be treated as a validation of the PPIs. A better approach then would be to detect both PIDMs and PPIs in the same article to derive meaningful data. We hypothesized that if an article contains the PIDM in its Methodology section, it is certain that PPIs were discovered in that study. A regular expressions (RegEx) based methodology was developed to classify the PubMed articles into one of the three PIDMs and then extract PPIs. This method was able to achieve an overall specificity of 83.6 and sensitivity of 78.2 in classification. The details of this method can be obtained from our published work [282]. Although we developed an efficient methodology for extraction of PPIs and PIDMs from published literature, it was not implemented into the present approach. In future, we plan to extend the current approach to include this work.

Limb regeneration data (proteomics and literature-derived) was mapped to the human orthologs since the genomic data for axolotl was not available. Once the genomic data becomes available, the same approach can be used to understand axolotl and *Xenopus* specific subnetworks. However, since the final goal is to identify targets which can confer limb regeneration ability in humans so that the soldiers who lose their limbs in wars and people who suffer amputations in accidents can be helped, we believe that the use of human orthologs will help achieve that objective.

Significance

Our study provides an exhaustive systems biology approach to compare regeneration competent and deficient subnetworks to show how the same proteins differentially inter-connect to confer regeneration-competence in axolotls. This approach also provides an in silico methodology to identify proteins that are not detected by experimental methods such as proteomics. Systems biology has the potential to map out numerous differential subnetworks that are crucial to blastema formation in regeneration-competent limbs and compare them to the pathways that characterize regeneration-deficient limbs, and to identify stem cell markers in regeneration. Humans are not able to regenerate appendages, nor can we regenerate skin, muscle, bone, or nerve across large gaps in these tissues. This approach will be a step forward in helping confer regenerative capacity on non-regenerating human tissues in future. The knowledge gained from the distinguishing features of limb regeneration at the systems level in amphibians can be used to chemically induce regeneration in mammalian systems. We believe that this research identified regeneration-promoting molecules which will fuel the research for regenerative medicine therapies.

Although this approach was implemented on limb regeneration, it is scalable and adaptable to compare any two given biological conditions. It provides novel intuitions that can further the understanding of the pathophysiological processes of the biological conditions being investigated and help predict the potential targets that can enhance drug discovery. Our findings show that although the proteins might be common between the two given biological conditions, they can have a high dissimilarity based on their biological and topological properties in the subnetwork. Hence, the discovery of differential subnetworks will also benefit the drug-repositioning pipeline since differential subnetworks can be easily eliminated as the potential targets for existing drugs.

Appendix

A1. Description of programs

#	Program Name	Input	Output
1	BioPortalRecommenderXML.java	XML of articles	Top5 ontologies for the articles supplied (printed in console)
2	BioPortalAnnotatorConceptListXML.java	XML of articles; list of ontologies to use (generated in program 1)	Unique List of Concept List Terms (XLS file)
3	PMIDAbstractsGenerateXLSFromXMLFile.java	XML of articles	XLS of the articles with PMIDs in first column followed by abstracts in second column
4	ConceptListMatchAbstractsXLS.java	XLS sheet of PMIDs and abstracts (generated in program 3); Concept list.xls (generated in program 2)	Two XLS sheets: one with weight (and all the factors needed to generate weight) for each matching term in every PMID; overall weight for each PMID
5	EvaluationMetric.java	XLS sheet containing overall PMID weights and the file containing known results	XLS file with values for all evaluation metrics mentioned in the methodology section with the user specified threshold interval

6	ExtractProteinsFromAbstracts.java	List of abstracts from database table: selected_abstracts	List of proteins stored in database table: protein_list
7	NetworkConstructionAxo.java	Database tables: axo_proteomics, biogrid_human_interactions_symbols, protein_list_frequency	Database tables: axo_present, axo_not_present, axo_combined
8	NetworkConstructionXeno.java	same as above, replace axo for xeno in the above table names	same as above, replace axo for xeno in the above table names
9	SubnetworkIdentification.java	Database tables: axo_combined, axo_proteomics, xeno_combined, xeno_proteomics, pmid_weight, pmid_protein, hprd, gene_go_bp, gene_kegg	XLS files for GO terms for the genes on subnetwork, subnetwork interactions, KEGG pathways for the proteins on subnetwork, seed nodes : proteins for the selected molecular class with their expression or literature weight wherever applicable
10	SubnetworkPValue.java	uses the XLS files generated in program 9	P-value associated with each subnetwork

11	SubnetworkGoPValue.java	uses the XLS files generated in program 9	Adds another column for p-values in the previously generated files containing GO terms
12	SubnetworkKeggPValue.java	uses the XLS files generated in program 9	Adds another column for p-values in the previously generated files containing KEGG pathways
13	DifferentialSubnetworks.java	uses the XLS files generated in program 9	Compares the axolotl and *Xenopus* subnetworks to identify differential subnetworks. Disco.xls contains the dissimilarity score for each common node in the subnetwork comparison
14	DiscoReport1.java	Disco.xls generated in program 13	Generates overall dissimilarity score for each subnetwork in axolotl
15	DiscoReport2.java	Disco.xls generated in program 13	Generates overall dissimilarity score for each subnetwork comparison

A2. Description of database tables

Table Name	**Column Name**	**Description**
abstract_data	pmid abstract_text	PMID and text of abstracts
axo_combined	gene_a gene_b status	Interaction data from proteomics and literature. “gene_a” contains the source interactor and “gene_b” contains the target interactor (gene symbols). “status” contains

		information about the source of protein. If both interacting proteins are derived from proteomics, status is “present” else “not_present”
axo_not_present	not_present_gene, gene_a gene_b	Proteins from literature and their interactions with proteomics data. “not_present_gene” contains the gene symbol for the literature-derived protein. gene_a and gene_b same as described above
axo_present	gene_a gene_b	Interactions where both proteins are from proteomics data. gene_a and gene_b same as described above
axolotl_proteomics_data	gene_name 1dpa 4dpa 7dpa	Proteomics data for axolotl. gene_name contains the gene symbol for the protein and 1dpa, 4dpa, and 7dpa contain respective fold change values
biogrid_human_interactions_symbols	bio_grid_id official_symbol_interactor_a official_symbol_interactor_b synonyms_interactor_a synonyms_interactor_b	Interactions and symbols for human data derived from BioGRID
protein_list	pmid protein_list	PMIDs and proteins extracted from literature with threshold of PMID weight > 2.5.
gene_go_bp	go_id gene_symbol description	GO BP for all the human proteins. “description” column contains the GO term description of GO IDs
gene_kegg	kegg_id	KEGG pathways for all the human

	gene_symbol description	proteins. “description” column contains the GO term description of GO IDs
hprd	title alt_title gene_symbol molecule_class molecular_function biological_process	HPRD data extracted from XML file
pmid_protein	pmid gene_name	PMID and gene symbol for protein extracted from literature
pmid_weight	pmid weight	Overall weight of each PMID extracted from literature
protein_list_frequency	protein count	Number of PMIDs in which a protein is present and gene symbol of the protein
selected_abstracts	pmid abstract_text	Abstracts for PMIDs with weight >2.5
uniprot_gene_names_human	entry entry_name status protein_names gene_names organism gene_names_primary gene_names_synonym	UniProt data (human and reviewed) extracted from XLS file downloaded from uniprot website
xeno_combined		*Xenopus* data, similar to axo_combined
xeno_not_present		*Xenopus* data, similar to xeno_not_present
xeno_present		*Xenopus* data, similar to axo_present
xenopus_proteomics_data	gene_name, 1dpa, 5dpa, 7dpa, 12dpa	*Xenopus* data, similar to axo_proteomics_data

References

1. Ideker T, Galitski T, Hood L: **A new approach to decoding life: systems biology**. *Annu Rev Genomics Hum Genet* 2001, **2**:343-372.
2. Ahn AC, Tewari M, Poon CS, Phillips RS: **The limits of reductionism in medicine: could systems biology offer an alternative?** *PLoS Med* 2006, **3**(6):e208.
3. Chuang HY, Hofree M, Ideker T: **A decade of systems biology**. *Annu Rev Cell Dev Biol* 2010, **26**:721-744.
4. Kitano H: **Systems biology: a brief overview**. *Science* 2002, **295**(5560):1662-1664.
5. **Gene Expression Omnibus** [http://www.ncbi.nlm.nih.gov/geo/]
6. Jhamb D, Rao N, Milner DJ, Song F, Cameron JA, Stocum DL, Palakal MJ: **Network based transcription factor analysis of regenerating axolotl limbs**. *BMC Bioinformatics* 2011, **12**:80.
7. Rao N, Jhamb D, Milner DJ, Li B, Song F, Wang M, Voss SR, Palakal M, King MW, Saranjami B *et al*: **Proteomic analysis of blastema formation in regenerating axolotl limbs**. *BMC Biol* 2009, **7**:83.
8. Boyack KW, Newman D, Duhon RJ, Klavans R, Patek M, Biberstine JR, Schijvenaars B, Skupin A, Ma N, Borner K: **Clustering more than two million biomedical publications: comparing the accuracies of nine text-based similarity approaches**. *PLoS One* 2011, **6**(3):e18029.
9. Lu Z: **PubMed and beyond: a survey of web tools for searching biomedical literature**. *Database (Oxford)* 2011, **2011**:baq036.
10. **Medical Subject Headings** [http://www.nlm.nih.gov/mesh/meshhome.html]
11. Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA, Gerstein M: **Genomic analysis of regulatory network dynamics reveals large topological changes**. *Nature* 2004, **431**(7006):308-312.
12. **IPA: Ingenuity Pathway Analysis** [http://www.ingenuity.com/products/ipa]
13. Ekins S, Nikolsky Y, Bugrim A, Kirillov E, Nikolskaya T: **Pathway mapping tools for analysis of high content data**. *Methods Mol Biol* 2007, **356**:319-350.
14. **What Is Text Mining?** [http://people.ischool.berkeley.edu/~hearst/text-mining.html]
15. Stevens R, Wroe C, Lord P, Goble C: **Ontologies in bioinformatics**. In: *Handbook on Ontologies*. Springer; 2003: 635-657.
16. Schuurman N, Leszczynski A: **Ontologies for bioinformatics**. *Bioinform Biol Insights* 2008, **2**:187-200.
17. **Fact Sheet PubMed®: MEDLINE® Retrieval on the World Wide Web** [http://www.nlm.nih.gov/pubs/factsheets/pubmed.html]
18. **National Center for Biotechnology Information** [http://www.ncbi.nlm.nih.gov/]
19. **European Bioinformatics Institute** [http://www.ebi.ac.uk/]
20. **UniProt** [http://www.uniprot.org/]
21. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A *et al*: **Human Protein Reference Database--2009 update**. *Nucleic Acids Res* 2009, **37**(Database issue):D767-772.

22. Chatr-Aryamontri A, Breitkreutz BJ, Heinicke S, Boucher L, Winter A, Stark C, Nixon J, Ramage L, Kolas N, O'Donnell L *et al*: **The BioGRID interaction database: 2013 update**. *Nucleic Acids Res* 2013, **41**(Database issue):D816-823.
23. Islamaj Dogan R, Murray GC, Neveol A, Lu Z: **Understanding PubMed user search behavior through log analysis**. *Database (Oxford)* 2009, **2009**:bap018.
24. **Automatic Term Mapping, PubMed** [http://www.nlm.nih.gov/bsd/disted/pubmedtutorial/020_040.html]
25. Bacchin M, Melucci M: **Symbol-Based Query Expansion Experiments at TREC 2005 Genomics Track**. In: *Proceedings of the Fourteenth Text Retrieval Conference: 2005*.
26. Hersh W, Price S, Donohoe L: **Assessing thesaurus-based query expansion using the UMLS Metathesaurus**. *Proc AMIA Symp* 2000:344-348.
27. Lu Z, Kim W, Wilbur WJ: **Evaluation of Query Expansion Using MeSH in PubMed**. *Inf Retr Boston* 2009, **12**(1):69-80.
28. Salton G, Buckley C: **Term-weighting approaches in automatic text retrieval**. *Information Processing & Management* 1988, **24**(5):513-523.
29. Belkin NJ, Kantor P, Fox EA, Shaw JA: **Combining the evidence of multiple query representations for information retrieval**. In: *TREC-2 Proceedings of the second conference on Text retrieval conference: 1995*.
30. Harman DK, Voorhees EM: **TREC: An overview**. *Annual Review of Information Science and Technology* 2006, **40**(1):113-155.
31. Hersh W, Voorhees E: **TREC genomics special issue overview**. *Information Retrieval* 2009, **12**(1):1-15.
32. Robertson SE, Walker S: **Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval**. In: *SIGIR '94 Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval: 1994*. 232 - 241.
33. Fontaine JF, Barbosa-Silva A, Schaefer M, Huska MR, Muro EM, Andrade-Navarro MA: **MedlineRanker: flexible ranking of biomedical literature**. *Nucleic Acids Res* 2009, **37**(Web Server issue):W141-146.
34. Smalheiser NR, Zhou W, Torvik VI: **Anne O'Tate: A tool to support user-driven summarization, drill-down and browsing of PubMed search results**. *J Biomed Discov Collab* 2008, **3**:2.
35. Yamamoto Y, Takagi T: **Biomedical knowledge navigation by literature clustering**. *J Biomed Inform* 2007, **40**(2):114-130.
36. Doms A, Schroeder M: **GoPubMed: exploring PubMed with the Gene Ontology**. *Nucleic Acids Res* 2005, **33**(Web Server issue):W783-786.
37. Perez-Iratxeta C, Bork P, Andrade MA: **XplorMed: a tool for exploring MEDLINE abstracts**. *Trends Biochem Sci* 2001, **26**(9):573-575.
38. Rebholz-Schuhmann D, Kirsch H, Arregui M, Gaudan S, Riethoven M, Stoehr P: **EBIMed--text crunching to gather facts for proteins from Medline**. *Bioinformatics* 2007, **23**(2):e237-244.
39. Fernandez JM, Hoffmann R, Valencia A: **iHOP web services**. *Nucleic Acids Res* 2007, **35**(Web Server issue):W21-26.

40. Tsai RT, Dai HJ, Lai PT, Huang CH: **PubMed-EX: a web browser extension to enhance PubMed search with text mining features**. *Bioinformatics* 2009, **25**(22):3031-3032.

41. Plake C, Schiemann T, Pankalla M, Hakenberg J, Leser U: **AliBaba: PubMed as a graph**. *Bioinformatics* 2006, **22**(19):2444-2445.

42. Douglas SM, Montelione GT, Gerstein M: **PubNet: a flexible system for visualizing literature derived networks**. *Genome Biol* 2005, **6**(9):R80.

43. Yu H, Kim T, Oh J, Ko I, Kim S, Han WS: **Enabling multi-level relevance feedback on PubMed by integrating rank learning into DBMS**. *BMC Bioinformatics* 2010, **11 Suppl 2**:S6.

44. States DJ, Ade AS, Wright ZC, Bookvich AV, Athey BD: **MiSearch adaptive pubMed search tool**. *Bioinformatics* 2009, **25**(7):974-976.

45. Yang Y, Pedersen JO: **A Comparative Study on Feature Selection in Text Categorization**. In: *ICML '97 Proceedings of the Fourteenth International Conference on Machine Learning: 1997*. Morgan Kaufmann Publishers Inc.: 412-420

46. Tzeras K, Hartmann S: **Automatic indexing based on Bayesian inference networks**. In: *SIGIR '93 Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval: 1993*. 22-35.

47. Wiener E, Pedersen JO, Weigend AS: **A Neural Network Approach to Topic Spotting**. In: *Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval: 1995*.

48. Yang Y: **Expert network: effective and efficient learning from human decisions in text categorization and retrieval**. In: *SIGIR '94 Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval: 1994*. 13-22.

49. Lewis DD, Ringuette M: **A Comparison of Two Learning Algorithms for Text Categorization**. In: *Third Annual Symposium on Document Analysis and Information Retrieval: 1994*.

50. Iliopoulos I, Enright AJ, Ouzounis CA: **Textquest: document clustering of Medline abstracts for concept discovery in molecular biology**. *Pac Symp Biocomput* 2001:384-395.

51. Papanikolaou N, Pavlopoulos GA, Pafilis E, Theodosiou T, Schneider R, Satagopam VP, Ouzounis CA, Eliopoulos AG, Promponas VJ, Iliopoulos I: **BioTextQuest+: a knowledge integration platform for literature mining and concept discovery**. *Bioinformatics* 2014.

52. Papanikolaou N, Pafilis E, Nikolaou S, Ouzounis CA, Iliopoulos I, Promponas VJ: **BioTextQuest: a web-based biomedical text mining suite for concept discovery**. *Bioinformatics* 2011, **27**(23):3327-3328.

53. **BioTextQuest+, A Biomedical Text Mining Suite for Concept Discovery** [http://bioinformatics.med.uoc.gr/cgi-bin/biotextquest/textQuest.cgi]

54. Yoo I, Hu X, Song IY: **A coherent graph-based semantic clustering and summarization approach for biomedical literature and a new summarization evaluation method**. *BMC Bioinformatics* 2007, **8 Suppl 9**:S4.

55. Yoo I, Hu X, Song IY: **Biomedical ontology improves biomedical literature clustering performance: a comparison study**. *Int J Bioinform Res Appl* 2007, **3**(3):414-428.

56. He D, Wu X: **Ontology-Based Feature Weighting for Biomedical Literature Classification**. In: *2006 IEEE International Conference on Information Reuse and Integration*. IEEE 2006: 280-285.

57. Hassanpour S, Das AK: **Ontology Based Text Mining of Concept Definitions in Biomedical Literature**. In: *Proceedings of the 3rd Canadian Semantic Web Symposium (CSWS2011): 2011*. 40-45.

58. Lin J, Wilbur WJ: **PubMed related articles: a probabilistic topic-based model for content similarity**. *BMC Bioinformatics* 2007, **8**:423.

59. Spärck Jones K, Walker S, Robertson SE: **A probabilistic model of information retrieval: development and comparative experiments**. *Information Processing & Management* 2000, **36**(6).

60. **Related Citations - PubMed** [http://www.nlm.nih.gov/bsd/disted/pubmedtutorial/020_190.html]

61. Consortium* TGO: **Gene Ontology: tool for the unification of biology**. *Nature Genetics* 2000, **25**(May).

62. **Unified Medical Language System** [http://www.nlm.nih.gov/research/umls/]

63. **Systematized Nomenclature of Medicine--Clinical Terms** [http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html]

64. Musen MA, Noy NF, Shah NH, Whetzel PL, Chute CG, Story MA, Smith B: **The National Center for Biomedical Ontology**. *J Am Med Inform Assoc* 2012, **19**(2):190-195.

65. Salvadores M, Alexander PR, Musen MA, Noy NF: **BioPortal as a Dataset of Linked Biomedical Ontologies and Terminologies in RDF**. *Semant Web* 2013, **4**(3):277-284.

66. Burdett T: **Zooma2 - A repository of annotation knowledge and curation API**. In: *International Society for Computational Biology*. 2013.

67. Jonquet C, Musen MA, Shah NH: **Building a biomedical ontology recommender web service**. *J Biomed Semantics* 2010, **1 Suppl 1**:S1.

68. Shah NH, Bhatia N, Jonquet C, Rubin D, Chiang AP, Musen MA: **Comparison of concept recognizers for building the Open Biomedical Annotator**. *BMC Bioinformatics* 2009, **10 Suppl 9**:S14.

69. Jonquet C, Lependu P, Falconer S, Coulet A, Noy NF, Musen MA, Shah NH: **NCBO Resource Index: Ontology-Based Search and Mining of Biomedical Resources**. *Web Semant* 2011, **9**(3):316-324.

70. Barabasi AL: **Scale-free networks: a decade and beyond**. *Science* 2009, **325**(5939):412-413.

71. Kitano H: **Computational systems biology**. *Nature* 2002, **420**(6912):206-210.

72. Barabasi AL: **Network medicine--from obesity to the "diseasome"**. *N Engl J Med* 2007, **357**(4):404-407.

73. Chautard E, Thierry-Mieg N, Ricard-Blum S: **Interaction networks: from protein functions to drug discovery. A review**. *Pathol Biol (Paris)* 2009, **57**(4):324-333.

74. Aderem A: **Systems biology: its practice and challenges**. *Cell* 2005, **121**(4):511-513.

75. Hood L, Perlmutter RM: **The impact of systems approaches on biological problems in drug discovery**. *Nat Biotechnol* 2004, **22**(10):1215-1217.

76. Barabasi AL, Oltvai ZN: **Network biology: understanding the cell's functional organization**. *Nat Rev Genet* 2004, **5**(2):101-113.

77. Barabasi AL, Albert R: **Emergence of scaling in random networks**. *Science* 1999, **286**(5439):509-512.

78. Killcoyne S, Carter GW, Smith J, Boyle J: **Cytoscape: a community-based framework for network modeling**. *Methods Mol Biol* 2009, **563**:219-239.

79. **Cyto-HUBBA** [http://hub.iis.sinica.edu.tw/cytoHubba/supplementary/index.htm]

80. Dezso Z, Nikolsky Y, Nikolskaya T, Miller J, Cherba D, Webb C, Bugrim A: **Identifying disease-specific genes based on their topological significance in protein networks**. *BMC Syst Biol* 2009, **3**:36.

81. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U: **Network motifs: simple building blocks of complex networks**. *Science* 2002, **298**(5594):824-827.

82. Shen-Orr SS, Milo R, Mangan S, Alon U: **Network motifs in the transcriptional regulation network of Escherichia coli**. *Nat Genet* 2002, **31**(1):64-68.

83. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I *et al*: **Transcriptional regulatory networks in Saccharomyces cerevisiae**. *Science* 2002, **298**(5594):799-804.

84. Setty Y, Mayo AE, Surette MG, Alon U: **Detailed map of a cis-regulatory input function**. *Proc Natl Acad Sci U S A* 2003, **100**(13):7702-7707.

85. Ratushny AV, Ramsey SA, Roda O, Wan Y, Smith JJ, Aitchison JD: **Control of transcriptional variability by overlapping feed-forward regulatory motifs**. *Biophys J* 2008, **95**(8):3715-3723.

86. Wuchty S, Oltvai ZN, Barabasi AL: **Evolutionary conservation of motif constituents in the yeast protein interaction network**. *Nat Genet* 2003, **35**(2):176-179.

87. Conant GC, Wagner A: **Convergent evolution of gene circuits**. *Nat Genet* 2003, **34**(3):264-266.

88. Hinman VF, Nguyen AT, Cameron RA, Davidson EH: **Developmental gene regulatory network architecture across 500 million years of echinoderm evolution**. *Proc Natl Acad Sci U S A* 2003, **100**(23):13356-13361.

89. Shoval O, Alon U: **SnapShot: network motifs**. *Cell* 2010, **143**(2):326-e321.

90. Alon U: **Network motifs: theory and experimental approaches**. *Nat Rev Genet* 2007, **8**(6):450-461.

91. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns**. *Proc Natl Acad Sci U S A* 1998, **95**(25):14863-14868.

92. Ihmels J, Friedlander G, Bergmann S, Sarig O, Ziv Y, Barkai N: **Revealing modular organization in the yeast transcriptional network**. *Nat Genet* 2002, **31**(4):370-377.

93. Tanay A, Sharan R, Shamir R: **Discovering statistically significant biclusters in gene expression data**. *Bioinformatics* 2002, **18 Suppl 1**:S136-144.

94. Hartwell LH, Hopfield JJ, Leibler S, Murray AW: **From molecular to modular cell biology**. *Nature* 1999, **402**(6761 Suppl):C47-52.

95. Ulitsky I, Shamir R: **Identification of functional modules using network topology and high-throughput data**. *BMC Syst Biol* 2007, **1**:8.

96. Maraziotis IA, Dimitrakopoulou K, Bezerianos A: **Growing functional modules from a seed protein via integration of protein interaction and gene expression data**. *BMC Bioinformatics* 2007, **8**:408.

97. Tornow S, Mewes HW: **Functional modules by relating protein interaction networks and gene expression**. *Nucleic Acids Res* 2003, **31**(21):6283-6289.

98. Dittrich MT, Klau GW, Rosenwald A, Dandekar T, Muller T: **Identifying functional modules in protein-protein interaction networks: an integrated exact approach**. *Bioinformatics* 2008, **24**(13):i223-231.

99. Wu H, Su Z, Mao F, Olman V, Xu Y: **Prediction of functional modules based on comparative genome analysis and Gene Ontology application**. *Nucleic Acids Res* 2005, **33**(9):2822-2837.

100. Wu Z, Zhao X, Chen L: **Identifying responsive functional modules from protein-protein interaction network**. *Mol Cells* 2009, **27**(3):271-277.

101. Singh R, Xu J, Berger B: **Global alignment of multiple protein interaction networks with application to functional orthology detection**. *Proc Natl Acad Sci U S A* 2008, **105**(35):12763-12768.

102. Kuchaiev O, Przulj N: **Integrative network alignment reveals large regions of global network similarity in yeast and human**. *Bioinformatics* 2011, **27**(10):1390-1396.

103. Wang K, Narayanan M, Zhong H, Tompa M, Schadt EE, Zhu J: **Meta-analysis of inter-species liver co-expression networks elucidates traits associated with common human diseases**. *PLoS Comput Biol* 2009, **5**(12):e1000616.

104. Zhang S, Ning XM, Ding C, Zhang XS: **Determining modular organization of protein interaction networks by maximizing modularity density**. *BMC Syst Biol* 2010, **4 Suppl 2**:S10.

105. Skinner J, Kotliarov Y, Varma S, Mine KL, Yambartsev A, Simon R, Huyen Y, Morgun A: **Construct and Compare Gene Coexpression Networks with DAPfinder and DAPview**. *BMC Bioinformatics* 2011, **12**:286.

106. Obayashi T, Nishida K, Kasahara K, Kinoshita K: **ATTED-II updates: condition-specific gene coexpression to extend coexpression analyses and applications to a broad range of flowering plants**. *Plant Cell Physiol* 2011, **52**(2):213-219.

107. Huang W, Cao X, Zhong S: **Network-based comparison of temporal gene expression patterns**. *Bioinformatics* 2010, **26**(23):2944-2951.

108. Diez D, Wheelock AM, Goto S, Haeggstrom JZ, Paulsson-Berne G, Hansson GK, Hedin U, Gabrielsen A, Wheelock CE: **The use of network analyses for elucidating mechanisms in cardiovascular disease**. *Mol Biosyst* 2010, **6**(2):289-304.

109. Liu CC, Chen WS, Lin CC, Liu HC, Chen HY, Yang PC, Chang PC, Chen JJ: **Topology-based cancer classification and related pathway mining using microarray data**. *Nucleic Acids Res* 2006, **34**(14):4069-4080.

110. Fuller TF, Ghazalpour A, Aten JE, Drake TA, Lusis AJ, Horvath S: **Weighted gene coexpression network analysis strategies applied to mouse weight**. *Mamm Genome* 2007, **18**(6-7):463-472.

111. Butte AJ, Tamayo P, Slonim D, Golub TR, Kohane IS: **Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks**. *Proc Natl Acad Sci U S A* 2000, **97**(22):12182-12186.

112. Steuer R, Kurths J, Fiehn O, Weckwerth W: **Observing and interpreting correlations in metabolomic networks**. *Bioinformatics* 2003, **19**(8):1019-1026.

113. Carter SL, Brechbuhler CM, Griffin M, Bond AT: **Gene co-expression network topology provides a framework for molecular characterization of cellular state**. *Bioinformatics* 2004, **20**(14):2242-2250.

114. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N: **Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data**. *Nat Genet* 2003, **34**(2):166-176.

115. Chuang HY, Lee E, Liu YT, Lee D, Ideker T: **Network-based classification of breast cancer metastasis**. *Mol Syst Biol* 2007, **3**:140.

116. Valavanis I, Spyrou G, Nikita K: **A similarity network approach for the analysis and comparison of protein sequence/structure sets**. *J Biomed Inform* 2010, **43**(2):257-267.

117. Hase T, Niimura Y, Tanaka H: **Difference in gene duplicability may explain the difference in overall structure of protein-protein interaction networks among eukaryotes**. *BMC Evol Biol* 2010, **10**:358.

118. Bell R, Hubbard A, Chettier R, Chen D, Miller JP, Kapahi P, Tarnopolsky M, Sahasrabuhde S, Melov S, Hughes RE: **A human protein interaction network shows conservation of aging processes between human and invertebrate species**. *PLoS Genet* 2009, **5**(3):e1000414.

119. Barrenas F, Chavali S, Holme P, Mobini R, Benson M: **Network properties of complex human disease genes identified through genome-wide association studies**. *PLoS One* 2009, **4**(11):e8090.

120. Ma HW, Zeng AP: **The connectivity structure, giant strong component and centrality of metabolic networks**. *Bioinformatics* 2003, **19**(11):1423-1430.

121. Guan Y, Myers CL, Lu R, Lemischka IR, Bult CJ, Troyanskaya OG: **A genomewide functional network for the laboratory mouse**. *PLoS Comput Biol* 2008, **4**(9):e1000165.

122. Gao J, Li Z: **Conserved network properties of helical membrane protein structures and its implication for improving membrane protein homology modeling at the twilight zone**. *J Comput Aided Mol Des* 2009, **23**(11):755-763.

123. Frenkel ZM, Trifonov EN, Snir S: **Structural relatedness via flow networks in protein sequence space**. *J Theor Biol* 2009, **260**(3):438-444.

124. Lappe M, Park J, Niggemann O, Holm L: **Generating protein interaction maps from incomplete data: application to fold assignment**. *Bioinformatics* 2001, **17 Suppl 1**:S149-156.

125. Almudevar A: **A hypothesis test for equality of bayesian network models**. *EURASIP J Bioinform Syst Biol* 2010, **2010**:947564.

126. Sharan R, Ideker T: **Modeling cellular machinery through biological network comparison**. *Nat Biotechnol* 2006, **24**(4):427-433.

127. Sharan R, Suthram S, Kelley RM, Kuhn T, McCuine S, Uetz P, Sittler T, Karp RM, Ideker T: **Conserved patterns of protein interaction in multiple species**. *Proc Natl Acad Sci U S A* 2005, **102**(6):1974-1979.

128. Kelley BP, Sharan R, Karp RM, Sittler T, Root DE, Stockwell BR, Ideker T: **Conserved pathways within bacteria and yeast as revealed by global protein network alignment**. *Proc Natl Acad Sci U S A* 2003, **100**(20):11394-11399.

129. Kalaev M, Smoot M, Ideker T, Sharan R: **NetworkBLAST: comparative analysis of protein networks**. *Bioinformatics* 2008, **24**(4):594-596.

130. Wuchty S, Almaas E: **Evolutionary cores of domain co-occurrence networks**. *BMC Evol Biol* 2005, **5**:24.

131. Wasserman WW, Palumbo M, Thompson W, Fickett JW, Lawrence CE: **Human-mouse genome comparisons to locate regulatory sites**. *Nat Genet* 2000, **26**(2):225-228.

132. Wang Y, Cui T, Zhang C, Yang M, Huang Y, Li W, Zhang L, Gao C, He Y, Li Y *et al*: **Global protein-protein interaction network in the human pathogen Mycobacterium tuberculosis H37Rv**. *J Proteome Res* 2010, **9**(12):6665-6677.

133. van Dijk AD, Morabito G, Fiers M, van Ham RC, Angenent GC, Immink RG: **Sequence motifs in MADS transcription factors responsible for specificity and diversification of protein-protein interaction**. *PLoS Comput Biol* 2010, **6**(11):e1001017.

134. Song N, Joseph JM, Davis GB, Durand D: **Sequence similarity network reveals common ancestry of multidomain proteins**. *PLoS Comput Biol* 2008, **4**(4):e1000063.

135. Plewczynski D, Rychlewski L, Ye Y, Jaroszewski L, Godzik A: **Integrated web service for improving alignment quality based on segments comparison**. *BMC Bioinformatics* 2004, **5**:98.

136. Mitra S, Gilbert JA, Field D, Huson DH: **Comparison of multiple metagenomes using phylogenetic networks based on ecological indices**. *ISME J* 2010, **4**(10):1236-1242.

137. Medini D, Covacci A, Donati C: **Protein homology network families reveal step-wise diversification of Type III and Type IV secretion systems**. *PLoS Comput Biol* 2006, **2**(12):e173.

138. Liang Z, Xu M, Teng M, Niu L: **Comparison of protein interaction networks reveals species conservation and divergence**. *BMC Bioinformatics* 2006, **7**:457.

139. Kuang R, Weston J, Noble WS, Leslie C: **Motif-based protein ranking by network propagation**. *Bioinformatics* 2005, **21**(19):3711-3718.

140. Fernandes LP, Annibale A, Kleinjung J, Coolen AC, Fraternali F: **Protein networks reveal detection bias and species consistency when analysed by information-theoretic methods**. *PLoS One* 2010, **5**(8):e12083.

141. Erten S, Li X, Bebek G, Li J, Koyuturk M: **Phylogenetic analysis of modularity in protein interaction networks**. *BMC Bioinformatics* 2009, **10**:333.

142. 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.

143. Flannick J, Novak A, Srinivasan BS, McAdams HH, Batzoglou S: **Graemlin: general and robust alignment of multiple large interaction networks**. *Genome Res* 2006, **16**(9):1169-1181.

144. Kelley BP, Yuan B, Lewitter F, Sharan R, Stockwell BR, Ideker T: **PathBLAST: a tool for alignment of protein interaction networks**. *Nucleic Acids Res* 2004, **32**(Web Server issue):W83-88.

145. Koyuturk M, Kim Y, Topkara U, Subramaniam S, Szpankowski W, Grama A: **Pairwise alignment of protein interaction networks**. *J Comput Biol* 2006, **13**(2):182-199.

146. Zhu D, Qin ZS: **Structural comparison of metabolic networks in selected single cell organisms**. *BMC Bioinformatics* 2005, **6**:8.

147. Mithani A, Preston GM, Hein J: **Rahnuma: hypergraph-based tool for metabolic pathway prediction and network comparison**. *Bioinformatics* 2009, **25**(14):1831-1832.

148. Mithani A, Hein J, Preston GM: **Comparative analysis of metabolic networks provides insight into the evolution of plant pathogenic and nonpathogenic lifestyles in Pseudomonas**. *Mol Biol Evol* 2011, **28**(1):483-499.

149. Mazurie A, Bonchev D, Schwikowski B, Buck GA: **Evolution of metabolic network organization**. *BMC Syst Biol* 2010, **4**:59.

150. Freilich S, Goldovsky L, Ouzounis CA, Thornton JM: **Metabolic innovations towards the human lineage**. *BMC Evol Biol* 2008, **8**:247.

151. Piovesan D, Martelli PL, Fariselli P, Zauli A, Rossi I, Casadio R: **BAR-PLUS: the Bologna Annotation Resource Plus for functional and structural annotation of protein sequences**. *Nucleic Acids Res* 2011, **39**(Web Server issue):W197-202.

152. Pesch R, Lysenko A, Hindle M, Hassani-Pak K, Thiele R, Rawlings C, Kohler J, Taubert J: **Graph-based sequence annotation using a data integration approach**. *J Integr Bioinform* 2008, **5**(2).

153. Janga SC, Diaz-Mejia JJ, Moreno-Hagelsieb G: **Network-based function prediction and interactomics: the case for metabolic enzymes**. *Metab Eng* 2011, **13**(1):1-10.

154. Hawkins T, Chitale M, Luban S, Kihara D: **PFP: Automated prediction of gene ontology functional annotations with confidence scores using protein sequence data**. *Proteins* 2009, **74**(3):566-582.

155. Gauthier JP, Legeai F, Zasadzinski A, Rispe C, Tagu D: **AphidBase: a database for aphid genomic resources**. *Bioinformatics* 2007, **23**(6):783-784.

156. Spallanzani L: **Concepts of generation and regeneration**. In: *A History of Regeneration Research* Edited by Dinsmore CE; 1991.

157. Stocum DL, Zupanc GK: **Stretching the limits: stem cells in regeneration science**. *Dev Dyn* 2008, **237**(12):3648-3671.

158. Stocum DL: **Regenerative Biology and Medicine**: Elsevier Inc.; 2006.

159. Goss RJ: **Tissue differentiation in regenerating antlers**. *Biol Deer Production* 1985 **22**:229-238.

160. Goss RJ: **Problems of antlerogesis**. *Clin Orthop Relat Res* 1970, **69**:227-238.

161. Illingworth CM: **Trapped fingers and amputated finger tips in children**. *J Pediatr Surg* 1974, **9**(6):853-858.

162. Borgens RB: **Mice regrow the tips of their foretoes**. *Science* 1982, **217**(4561):747-750.

163. Han M, Yang X, Farrington JE, Muneoka K: **Digit regeneration is regulated by Msx1 and BMP4 in fetal mice**. *Development* 2003, **130**(21):5123-5132.

164. Goss RJ, Grimes LN: **Tissue interactions in the regeneration of rabbit ear holes** *Am Zool* 1975, **12**:151-157.

165. Heber-Katz E, Leferovich JM, Bedelbaeva K, Gourevitch D: **Spallanzani's mouse: a model of restoration and regeneration**. *Curr Top Microbiol Immunol* 2004, **280**:165-189.

166. Bryant SV, Endo T, Gardiner DM: **Vertebrate limb regeneration and the origin of limb stem cells**. *Int J Dev Biol* 2002, **46**(7):887-896.

167. Nye HL, Cameron JA, Chernoff EA, Stocum DL: **Regeneration of the urodele limb: a review**. *Dev Dyn* 2003, **226**(2):280-294.

168. Morrison JI, Loof S, He P, Simon A: **Salamander limb regeneration involves the activation of a multipotent skeletal muscle satellite cell population**. *J Cell Biol* 2006, **172**(3):433-440.

169. Carlson BM: **Principles of Regenerative Biology**: Academic Press; 2007.

170. Iten LE, Bryant SV: **Forelimb regeneration from different levels of amputation in the newt Notophthalmus viridesces: Length, rate and stages** *W Roux Archiv* 1973 **173**:263-282.

171. Stocum DL: **Stages of forelimb regeneration in Ambystoma maculatum**. *J Exp Zool* 1979, **209**(3):395-416.

172. Mescher AL: **The cellular basis of limb regeneration in urodeles**. *Int J Dev Biol* 1996, **40**(4):785-795.

173. Chalkley DT: **A quantitative histological analysis of forelimb regeneration in Triturus viridescens** *J Morphol* 1954, **94**:21-70.

174. Chalkley DT: **The cellular basis of limb regeneration** In: *Regeneration in Vertebrates.* Edited by Thornton CS. Chicago: University of Chicago Press; 1956: pp 34-56.

175. Hay ED, Fischman DA: **Origin of the blastema in regenerating limbs of the newt Triturus viridescens. An autoradiographic study using tritiated thymidine to follow cell proliferation and migration**. *Dev Biol* 1961, **3**:26-59.

176. Kelly DJ, Tassava RA: **Cell division and ribonucleic acid synthesis during the initiation of limb regeneration in larval axolotls (Ambystoma mexicanum)**. *J Exp Zool* 1973, **185**(1):45-54.

177. Loyd RM, Tassava RA: **DNA synthesis and mitosis in adult newt limbs following amputation and insertion into the body cavity**. *J Exp Zool* 1980, **214**(1):61-69.

178. McCullough WD, Tassava RA: **Determination of the blastema cell cycle in regenerating limbs of the larval axolotl, Ambystoma mexicanum**. *Ohio J Sci* 1976 **76**:63-65.

179. Tassava RA, Goldhamer DJ, Tomlinson BL: **Cell cycle controls and the role of nerves and the regenerate epithelium in urodele forelimb regeneration: possible modifications of basic concepts**. *Biochem Cell Biol* 1987, **65**(8):739-749.

180. Kumar A, Godwin JW, Gates PB, Garza-Garcia AA, Brockes JP: **Molecular basis for the nerve dependence of limb regeneration in an adult vertebrate**. *Science* 2007, **318**(5851):772-777.

181. Steen TP: **Stability of chondrocyte differentiation and contribution of muscle to cartilage during limb regeneration in the axolotl (Siredon mexicanum)**. *J Exp Zool* 1968, **167**(1):49-78.

182. Cameron JA, Hinterberger TJ: **Regional differences in the distribution of of myogenic and chondrogenic cells in axolotl limb blastemas** *J Exp Zool* 1984 **232**:269-275.

183. Kragl M, Knapp D, Nacu E, Khattak S, Maden M, Epperlein HH, Tanaka EM: **Cells keep a memory of their tissue origin during axolotl limb regeneration**. *Nature* 2009, **460**(7251):60-65.

184. Sauer U, Heinemann M, Zamboni N: **Genetics. Getting closer to the whole picture**. *Science* 2007, **316**(5824):550-551.

185. Putta S, Smith JJ, Walker JA, Rondet M, Weisrock DW, Monaghan J, Samuels AK, Kump K, King DC, Maness NJ *et al*: **From biomedicine to natural history research: EST resources for ambystomatid salamanders**. *BMC Genomics* 2004, **5**(1):54.

186. Monaghan JR, Epp LG, Putta S, Page RB, Walker JA, Beachy CK, Zhu W, Pao GM, Verma IM, Hunter T *et al*: **Microarray and cDNA sequence analysis of transcription during nerve-dependent limb regeneration**. *BMC Biol* 2009, **7**:1.

187. Bodemer CW: **Distribution of ribonucleic acid in the urodele limb as determined by autoradiographic localization of uridine-H3.** . *Anat Rec* 1962 **142**:147-148.

188. Bodemer CW, Everett NB: **Localization of newly synthesized proteins in regenerating newt limbs as determined by radioautographic localization of injected methinine-S35.** . *Dev Biol* 1959, **1**:327-342.

189. Urbani E: **Proteolytic enzymes in regeneration**. In: *Regeneration in Animals.* Edited by Kiortsis V, Trampusch HAL. Amsterdam: North-Holland Pub Co; 1965: pp 39-55.

190. Anton HJ: **The origin of blastema cells and protein synthesis during forelimb regeneration in Triturus**. In: *Regeneration in Animals.* Edited by Kiortsis V, Trampusch HAL. Amsterdam: North-Holland Pub Co; 1965: pp 377-395.

191. Lebowitz P, Singer M: **Neurotrophic control of protein synthesis in the regenerating limb of the newt, Triturus**. *Nature* 1970, **225**(5235):824-827.

192. Singer M, Ilan J: **Nerve-dependent regulation of absolute rates of protein synthesis in newt limb regenerates. Measurement of methionine specific activity in peptidyl-tRNA of the growing polypeptide chain**. *Dev Biol* 1977, **57**(1):174-187.

193. Dearlove GE, Stocum DL: **Denervation-induced changes in soluble protein content during forelimb regeneration in the adult newt, Notophthalmus viridescens**. *J Exp Zool* 1974, **190**(3):317-328.

194. Slack JM: **Protein synthesis during limb regeneration in the axolotl**. *J Embryol Exp Morphol* 1982, **70**:241-260.

195. Tsonis PA: **A comparative two-dimensional gel protein database of the intact and regenerating newt limbs**. *Electrophoresis* 1993, **14**(1-2):148-156.

196. Tsonis PA, Mescher AL, Del Rio-Tsonis K: **Protein synthesis in the newt regenerating limb. Comparative two-dimensional PAGE, computer analysis and protein sequencing**. *Biochem J* 1992, **281 (Pt 3)**:665-668.
197. Stocum DL, Rao N: **Mechanisms of Blastema Formation in Regenerating Amphibian Limbs**. In: *Principles of Regenerative Medicine.* Edited by Lanza R, Thompson J, Nerem R: Elsevier/ Academic Press San Diego; in press.
198. Dent J: **Limb regeneration in larvae and metamorphosing individuals of the South African clawed toad**. *J Morph* 1962, **110**:61 - 77.
199. Suzuki M, Yakushiji N, Nakada Y, Satoh A, Ide H, Tamura K: **Limb regeneration in Xenopus laevis froglet**. *TSW Develop Embryol* 2006, **1**(S1):26 - 37.
200. Kawasuki A, Sagawa N, Hayashi S, Yokoyama H, Tamura K: **Wound healing in mammals and amphibians: toward limb regeneration in mammals**. *Curr Topics Microbio Immunol* 2013, **367**:33 - 74.
201. Wolfe A, Nye H, Cameron J: **Extent of ossification at the amputation plane is correlated with the decline of blastema formation and regeneration in Xenopus laevis hindlimbs**. *Dev Dyn* 2000, **218**:681 - 697.
202. Sessions S, Bryant S: **Evidence that regenerative ability is an intrinsic property of limb cells in Xenopus**. *J Exp Zool* 1988, **247**:39 - 44.
203. Filoni S, Velloso C, Bernardini S, Cannata S: **Acquisition of nerve dependence for the formation of a regeneration blastema in amputated hindlimbs of larval Xenopus laevis: the role of limb innervation and that of limb differentiation**. *J Exp Zool* 1995, **1995**(273):327 - 341.
204. King MW, Nguyen T, Calley J, Harty MW, Muzinich MC, Mescher AL, Chalfant C, N'Cho M, McLeaster K, McEntire J *et al*: **Identification of genes expressed during Xenopus laevis limb regeneration by using subtractive hybridization**. *Dev Dyn* 2003, **226**(2):398-409.
205. Grow M, Neff AW, Mescher AL, King MW: **Global analysis of gene expression in Xenopus hindlimbs during stage-dependent complete and incomplete regeneration**. *Dev Dyn* 2006, **235**(10):2667-2685.
206. King MW, Neff AW, Mescher AL: **Proteomics analysis of regenerating amphibian limbs: changes during the onset of regeneration**. *Int J Dev Biol* 2009, **53**(7):955-969.
207. Skowron S, Komala Z: **Limb regeneration in postmetamorphic Xenopus laevis**. *Folia Biol Krakow* 1957, **5**:53 - 72.
208. Khan P, Liversage R: **Ultrastructural comparison between regenerating and developing hindlimbs of Xenopus laevis tadpoles**. *Growth Develop Aging* 1990, **54**:173 - 181.
209. Goss R, Holt R: **Epimorphic vs. tissue regeneration in Xenopus forelimbs**. *J Exp Zool* 1992, **261**:451 - 457.
210. Suzuki M, Satoh A, Ide H, Tamura K: **Nerve-dependent and -independent events in blastema formation during Xenopus froglet limb regeneration**. *Dev Biol* 2005, **286**:361 - 375.
211. Suzuki M, Satoh A, Ide H, Tamura K: **Transgenic Xenopus with prx1 limb enhancer reveals crucial contribution of MEK/ERK and PI3K/AKT**

pathways in blastema formation during limb regeneration. *Dev Biol* 2007, **2007**(304):675 - 686.

212. Satoh A, James M, Gardiner D: **The role of nerve signaling in limb genesis and agenesis during axolotl limb regeneration**. *J Bone Joint Surg* 2009, **91**(S4):90 - 98.

213. Furlong S, Heidemann M, Bromley S: **Fine structure of the forelimb regenerate of the African clawed toad, Xenopus laevis**. *Anat Rec* 1985, **1985**(211):444 - 449.

214. Korneluk R, Liversage R: **Effects of radius-ulna removal on forelimb regeneration in Xenopus laevis froglets**. *J Embryol Exp Morph* 1984, **82**:9 - 24.

215. Komala Z: **Poro' wnawcze badania nad przebiegiem ontogenezy I regen eracji konczynkon'czyn kijanek Xenopus laevis w ro' znychro' znych okresach rozwojowych**. *Folia Biol Krakow* 1957, **5**:1 - 52.

216. Mescher A, Neff A: **Limb regeneration in amphibians: immunological considerations**. *TheScientificWorldJOURNAL* 2006, **6**(Suppl 1):1 - 11.

217. Mescher A, Neff A: **Regenerative capacity and the developing immune system**. *Adv in Biochem Eng/Biotechnol* 2005, **93**:39 - 66.

218. Harty M, Neff A, King M, Mescher A: **Regeneration or scarring: an immunologic perspective**. *Devel Dynam* 2003, **226**:268 - 279.

219. Pearl E, Barker R, Day R, Beck C: **Identification of genes associated with regenerative success of Xenopus laevis hindlimbs**. *BMC Dev Biol* 2008, **8**:66.

220. King M, Nguyen T, Calley J, Harty M, Muzinich M, Mescher A, Chalfant C, N'Cho M, McLeaster K, McEntire J *et al*: **Identification of genes expressed during Xenopus laevis limb regeneration by using subtractive hybridization**. *Develop Dyn* 2003, **226**:398 - 409.

221. Grow M, Neff A, Mescher A, King M: **Global analysis of gene expression in Xenopus hindlimbs during stage-dependent complete and incomplete regeneration**. *Dev Dyn* 2006, **235**:2667 - 2685.

222. Endo T, Tamura K, Ide H: **Analysis of gene expressions during Xenopus forelimb regeneration**. *Dev Biol* 2000, **220**:296 - 306.

223. Ohgo S, Itoh A, Suzuki M, Satoh A, Yokoyama H, Tamura K: **Analysis of hoxa11 and hoxa13 expression during patternless limb regeneration in Xenopus**. *Dev Biol* 2010, **338**:148 - 157.

224. Yakushiji N, Suzuki M, Satoh A, Sagai T, Shiroishi T, Kobayashi H, Sasaki H, Ide H, Tamura K: **Correlation between Shh expression and DNA methylation status of the limb-specific Shh enhancer region during limb regeneration in amphibians**. *Dev Biol* 2007, **312**:171 - 182.

225. Fitzpatrick DPG, You JS, Bemis KG, Wery JP, Ludwig J, Wang M: **Searching for potential biomarkers of cisplatin resistance in human ovarian cancer using a label-free LC/MS-based protein quantification method**. *PROTEOMICS - Clin Appl* 2007, **1**:246-263.

226. Rao N, Song F, Jhamb D, Wang M, Milner D, Price N, Belecky-Adams T, Palakal M, Cameron J, Li B *et al*: **Proteomic analysis of fibroblastema formation in regenerating hind limbs of Xenopus laevis froglets and comparison to axolotl**. *BMC Developmental Biology* 2014, **14**(1):32.

227. Higgs RE, Knierman MD, Gelfanova V, Butler JP, Hale JE: **Comprehensive label-free method for the relative quantification of proteins from biological samples**. *J Proteome Res* 2005, **4**(4):1442-1450.
228. Limpert E, Stahel WA, Abbot M: **Log-normal distributions across the sciences: Keys and clues**. *Biosci* 2001 **51**:341-352.
229. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias**. *Bioinformatics* 2003, **19**(2):185-193.
230. Reiner A, Yekutieli D, Benjamini Y: **Identifying differentially expressed genes using false discovery rate controlling procedures**. *Bioinformatics* 2003, **19**(3):368-375.
231. Hakes L, Pinney JW, Robertson DL, Lovell SC: **Protein-protein interaction networks and biology--what's the connection?** *Nat Biotechnol* 2008, **26**(1):69-72.
232. Kerrien S, Orchard S, Montecchi-Palazzi L, Aranda B, Quinn AF, Vinod N, Bader GD, Xenarios I, Wojcik J, Sherman D *et al*: **Broadening the horizon--level 2.5 of the HUPO-PSI format for molecular interactions**. *BMC Biol* 2007, **5**:44.
233. Krallinger M, Vazquez M, Leitner F, Salgado D, Chatr-Aryamontri A, Winter A, Perfetto L, Briganti L, Licata L, Iannuccelli M *et al*: **The Protein-Protein Interaction tasks of BioCreative III: classification/ranking of articles and linking bio-ontology concepts to full text**. *BMC Bioinformatics* 2011, **12 Suppl 8**:S3.
234. Smith L, Rindflesch T, Wilbur WJ: **MedPost: a part-of-speech tagger for bioMedical text**. *Bioinformatics* 2004, **20**(14):2320-2321.
235. **LingPipe 4.1.0.** [http://alias-i.com/lingpipe]
236. Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, Tudorache T, Musen MA: **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):W541-545.
237. Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes**. *Nucleic Acids Res* 2000, **28**(1):27-30.
238. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M: **The KEGG resource for deciphering the genome**. *Nucleic Acids Res* 2004, **32**(Database issue):D277-280.
239. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA: **Circos: an information aesthetic for comparative genomics**. *Genome Res* 2009, **19**(9):1639-1645.
240. **The R Project for Statistical Computing** [http://www.r-project.org/]
241. **Gplots, a R package** [http://cran.r-project.org/web/packages/gplots/index.html]
242. Saldanha AJ: **Java Treeview--extensible visualization of microarray data**. *Bioinformatics* 2004, **20**(17):3246-3248.
243. Gilbert SF: **Developmental Biology**, 9 edn: Sinauer Associates Inc, Sunderland, MA; 2010.

244. Palakal M, Stephens M, Mukhopadhyay S, Raje R, Rhodes S: **A multi-level text mining method to extract biological relationships**. *Proc IEEE Comput Soc Bioinform Conf* 2002, **1**:97-108.

245. Chen X, Song F, Li J, Jhamb D, Alshalchi S, Hicks E, Bottino MC, Palakal MJ, Stocum DL: **The Axolotl Fibula as a Model to Induce Regeneration Across Large Segment Defects in Long Bones of the Extremities**. *Tissue Engineering A, submitted* 2015.

246. Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: **DAVID: Database for Annotation, Visualization, and Integrated Discovery**. *Genome Biol* 2003, **4**(5):P3.

247. **NCIT: National Cancer Institute Thesaurus** [http://bioportal.bioontology.org/ontologies/NCIT]

248. 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 Res* 2003, **13**(11):2498-2504.

249. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT *et al*: **Gene expression profiling predicts clinical outcome of breast cancer**. *Nature* 2002, **415**(6871):530-536.

250. Yook JI, Li XY, Ota I, Hu C, Kim HS, Kim NH, Cha SY, Ryu JK, Choi YJ, Kim J *et al*: **A Wnt-Axin2-GSK3beta cascade regulates Snail1 activity in breast cancer cells**. *Nat Cell Biol* 2006, **8**(12):1398-1406.

251. Lee KH, Choi EY, Kim MK, Lee SH, Jang BI, Kim TN, Kim SW, Song SK, Kim JR, Jung BC: **Hepatoma-derived growth factor regulates the bad-mediated apoptotic pathway and induction of vascular endothelial growth factor in stomach cancer cells**. *Oncol Res* 2010, **19**(2):67-76.

252. Enomoto H, Nakamura H, Liu W, Yoshida K, Okuda Y, Imanishi H, Saito M, Shimomura S, Hada T, Nishiguchi S: **Hepatoma-derived growth factor is induced in liver regeneration**. *Hepatol Res* 2009, **39**(10):988-997.

253. Knoepfler PS: **Why myc? An unexpected ingredient in the stem cell cocktail**. *Cell Stem Cell* 2008, **2**(1):18-21.

254. Eilers M, Eisenman RN: **Myc's broad reach**. *Genes Dev* 2008, **22**(20):2755-2766.

255. Stojadinovic O, Brem H, Vouthounis C, Lee B, Fallon J, Stallcup M, Merchant A, Galiano RD, Tomic-Canic M: **Molecular pathogenesis of chronic wounds: the role of beta-catenin and c-myc in the inhibition of epithelialization and wound healing**. *Am J Pathol* 2005, **167**(1):59-69.

256. Hourdry J, Geraudie J, Singer M, Mechali M: **Expression of the c-Myc proto-oncogene in the ofrelimb regenerate of the newt *Notophthalmus Viridescens*, visualized by in situ hybridization**. In: *M Singer Symposium*. 1988: 307-313.

257. Geraudie J, Hourdry J, Boehm K, Singer M, Mechali M: **c-Myc proto-oncogene expression during newt limb regeneration**. In: *Recent Trends in Regeneration Research.* Edited by Kiortsis V, Koussoulallos S, Wallace H, vol. 172. New York: Plenum Press; 1989: 27-36.

258. Maki N, Suetsugu-Maki R, Tarui H, Agata K, Del Rio-Tsonis K, Tsonis PA: **Expression of stem cell pluripotency factors during regeneration in newts**. *Dev Dyn* 2009, **238**(6):1613-1616.

259. Wierstra I: **Sp1: emerging roles--beyond constitutive activation of TATA-less housekeeping genes**. *Biochem Biophys Res Commun* 2008, **372**(1):1-13.

260. Tan NY, Khachigian LM: **Sp1 phosphorylation and its regulation of gene transcription**. *Mol Cell Biol* 2009, **29**(10):2483-2488.

261. Safe S, Abdelrahim M: **Sp transcription factor family and its role in cancer**. *Eur J Cancer* 2005, **41**(16):2438-2448.

262. Cadinouche MZ, Liversage RA, Muller W, Tsilfidis C: **Molecular cloning of the Notophthalmus viridescens radical fringe cDNA and characterization of its expression during forelimb development and adult forelimb regeneration**. *Dev Dyn* 1999, **214**(3):259-268.

263. Crews L, Gates PB, Brown R, Joliot A, Foley C, Brockes JP, Gann AA: **Expression and activity of the newt Msx-1 gene in relation to limb regeneration**. *Proc Biol Sci* 1995, **259**(1355):161-171.

264. Shimizu-Nishikawa K, Tsuji S, Yoshizato K: **Identification and characterization of newt rad (ras associated with diabetes), a gene specifically expressed in regenerating limb muscle**. *Dev Dyn* 2001, **220**(1):74-86.

265. Koshiba K, Kuroiwa A, Yamamoto H, Tamura K, Ide H: **Expression of Msx genes in regenerating and developing limbs of axolotl**. *J Exp Zool* 1998, **282**(6):703-714.

266. Schnapp E, Tanaka EM: **Quantitative evaluation of morpholino-mediated protein knockdown of GFP, MSX1, and PAX7 during tail regeneration in Ambystoma mexicanum**. *Dev Dyn* 2005, **232**(1):162-170.

267. Takahashi K, Yamanaka S: **Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors**. *Cell* 2006, **126**(4):663-676.

268. Yu J, Vodyanik MA, Smuga-Otto K, Antosiewicz-Bourget J, Frane JL, Tian S, Nie J, Jonsdottir GA, Ruotti V, Stewart R *et al*: **Induced pluripotent stem cell lines derived from human somatic cells**. *Science* 2007, **318**(5858):1917-1920.

269. Sharma M, Naslavsky N, Caplan S: **A role for EHD4 in the regulation of early endosomal transport**. *Traffic* 2008, **9**(6):995-1018.

270. Kuo HJ, Tran NT, Clary SA, Morris NP, Glanville RW: **Characterization of EHD4, an EH domain-containing protein expressed in the extracellular matrix**. *J Biol Chem* 2001, **276**(46):43103-43110.

271. Stocum DL: **Wound Repair, Regeneration and Artificial Tissues**. Austin, TX: RG Landes Co; 1995

272. Mescher AL, Neff AW: **Regenerative capacity and the developing immune system**. *Adv Biochem Eng Biotechnol* 2005, **93**:39-66.

273. Mescher AL, Neff AW: **Limb regeneration in amphibians: immunological considerations**. *ScientificWorldJournal* 2006, **6 Suppl 1**:1-11.

274. Caldwell RL, Caprioli RM: **Tissue profiling by mass spectrometry: a review of methodology and applications**. *Mol Cell Proteomics* 2005, **4**(4):394-401.

275. Caldwell RL, Opalenik SR, Davidson JM, Caprioli RM, Nanney LB: **Tissue profiling MALDI mass spectrometry reveals prominent calcium-binding proteins in the proteome of regenerative MRL mouse wounds**. *Wound Repair Regen* 2008, **16**(3):442-449.

276. Desrichard A, Bidet Y, Uhrhammer N, Bignon YJ: **CHEK2 contribution to hereditary breast cancer in non-BRCA families**. *Breast Cancer Res* 2011, **13**(6).

277. Choo YS, Vogler G, Wang DL, Kalvakuri S, Iliuk A, Tao WA, Bodmer R, Zhang ZH: **Regulation of parkin and PINK1 by neddylation**. *Human Molecular Genetics* 2012, **21**(11):2514-2523.

278. Parsons DW, Li M, Zhang X, Jones S, Leary RJ, Lin JC, Boca SM, Carter H, Samayoa J, Bettegowda C *et al*: **The genetic landscape of the childhood cancer medulloblastoma**. *Science* 2011, **331**(6016):435-439.

279. Nagamani SC, Erez A, Bay C, Pettigrew A, Lalani SR, Herman K, Graham BH, Nowaczyk MJ, Proud M, Craigen WJ *et al*: **Delineation of a deletion region critical for corpus callosal abnormalities in chromosome 1q43-q44**. *Eur J Hum Genet* 2012, **20**(2):176-179.

280. Jhamb D, Pradhan MP, Duraiswamy P, Desai A, Palakal MJ: **A systems biology framework for the downstream analysis of the whole genome sequencing data**. In: *IEEE International Conference on Computational Advances in Bio and Medical Sciences*. 2014.

281. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M: **BioGRID: a general repository for interaction datasets**. *Nucleic Acids Res* 2006, **34**(Database issue):D535-539.

282. Jhamb D, Krishnan A, Pandit Y, Duraiswamy P, Palakal M, Palakal MJ: **Protein interaction detection method to classify the documents from biomedical literature and obtain relevant protein-protein interactions**. In: *IEEE International Conference on Computational Advances in Bio and Medical Sciences*. 2014.

CURRICULUM VITAE

Deepali Jhamb

Education

Doctor of Philosophy, Informatics, 2015: Indiana University, Indianapolis, IN
Masters of Science, Life Sciences, 2008: Indiana State University, Terre Haute, IN
Masters of Science, Medical Biochemistry, 2006: Manipal University, Karnataka, India
Bachelors of Science, Microbiology (Hons.), 2003: Delhi University, New Delhi, India

Work Experience

Scientific Investigator, GlaxoSmithKline, King of Prussia, PA , August 2014 to Present

- Computational biology analysis of clinical and preclinical data for target discovery and validation

Research Assistant, IUPUI, IN, May 2008 to July 2014

- Segment Defect Regeneration: To identify growth factors involved in segment defect regeneration across a critical size defect in Axolotl limbs. Nine growth factors were identified using text mining, systems biology and statistical analysis. These growth factors were further tested in these animals using scaffolds and have been shown to regenerate muscle and cartilage in the previously non-regenerating defects.
- Whole Genome Sequencing Project: Korean Personal Genome Project data provided by CAMDA, 2013 (as a part of the ISMB conference) was used to study the SNP profile. This data was compared with the 1000 genome data using systems biology approaches to identify significant variants. Overall, we found substantial evidence for the high number of variants related to the neurodegenerative disorders and tumors in the KPGP dataset.
- Proteomics Data Analysis: LC/MS/MS datasets at different time points were analyzed for two different biological models – limb regeneration-competent Axolotl and limb regeneration-deficient Xenopus. These datasets were analyzed using differential network analysis to identify previously unknown key targets in limb regeneration.
- Algorithm Design and Development: Text Mining algorithms were developed for condition-specific data extraction from biomedical literature. Systems biology methods were designed and applied to identify differential subnetworks and key targets between two given biological systems.
- Database Design and Development: Limb regeneration database was created to help biology scientists with easy access to the data and discover new patterns.
- Designed and mentored the bioinformatics research projects of undergraduate students and SEED high school students.

Research Assistant, IUSM, Terre Haute, IN, August 2006 - May 2008

- Generation of transgenic Xenopus laevis expressing the limb regeneration-competent gene, SALL4. This involved the use of several molecular biology techniques.
- Proteomic data analysis of limb regeneration-competent vs. deficient *Xenopus laevis*.

Grants

Designed and wrote the bioinformatics strategy for the following grants:

- "Regenerative medicine for battlefield injuries". Dept. of Defense (DoD) Grant, 2011.
- "Decoding the code of limb regeneration - a systems biology approach”. W.M Keck Foundation grant, 2010.
- "Systems analysis of epimorphic and segment defect regeneration". MURI grant, 2008 (Was not approved but received good reviews).
- “Generate transgenic Xenopus expressing SALL4”. Graduate student research grant, ISU, 2007.

Computational Skills

- Operating System - UNIX, Windows-Vista/XP/7
- Languages - JAVA, PERL, Python
- Database - MySQL
- Statistical software/packages - R

Bioinformatics Skills

- NGS Analysis - SAMtools, BWA, Bowtie, vcftools, GATK, GALAXY, Picard, ANNOVAR, SnpEff, SIFT, PolyPhen-2, Ingenuity Variant Analysis, IGV, NCBI Genome Browser
- Microarray Analysis - GEO, Array express, Gene expression atlas, GSEA, SAM, EXPANDER
- Network Analysis - MetaCore, Ingenuity, NetworkBLAST, PathBLAST, SANDY, Netgrep, Motifsearch
- Network Visualization - Cytoscape, Circos, Gephi, NAViGaTOR
- Databases - GO, KEGG, HPRD, BioGRID, UniProt, GeneCards
- Others - UCSC genome browser, Connectivity Map, GenePattern, Experimental Factor Ontology

Publications and Poster Presentations

Publications

- Rao, N., Song, F., Jhamb, D., Price, N., Wang, M., Adams, T., Chen, X., Palakal, M., Milner, D., Cameron, J., Li, B., Stocum, D. (2014). “Proteomic Analysis of Fibroblastema Formation in Regenerating Hind Limbs of Xenopus Laevis Froglets and Comparison to Axolotl.” BMC Developmental Biology, 14:32
- Jhamb D, Pradhan MP, Duraiswamy P, Desai A, Palakal MJ: A systems biology framework for the downstream analysis of the whole genome sequencing data. In: IEEE International Conference on Computational Advances in Bio and Medical Sciences. 2014.
- Jhamb, D., Krishnan, A., Pandit, Y., Duraiswamy, P., Palakal, M.: Protein interaction detection method to classify the documents from biomedical literature and obtain relevant protein-protein interactions. In: IEEE International Conference on Computational Advances in Bio and Medical Sciences. 2014.
- Jhamb, D., N. Rao, Milner, D., Song, F., Cameron, J., Stocum, D., Palakal, M. (2011). "Network based transcription factor analysis of regenerating axolotl limbs."

BMC Bioinformatics 12: 80.

- Rao, N., Jhamb, D., Milner, D., Li, B., Song, F., Wang, M., Voss, R., Palakal, M., King, M., Saranjami, B., Nye, H., Cameron, J., Stocum, D. (2009). "Proteomic analysis of blastema formation in regenerating axolotl limbs." BMC Biol 7: 83.
- Chen, X., Song, F., Li, J., Jhamb D, Alshalchi, S., Hicks, E., Bottino, M., Palakal, M., Stocum, D.: The Axolotl Fibula as a Model to Induce Regeneration Across Large Segment Defects in Long Bones of the Extremities (submitted)
- Jhamb, D., Stocum, D., Palakal, M.: Condition-specific data mining and differential subnetwork analysis for limb regeneration. (in preparation)

Poster Presentations

- Awonusi, D., Jhamb, D., Palakal, M. (2011) "Assigning biological relevance to the signaling networks", Research Day, IUPUI.
- Jhamb, D., Rao, N., Milner, D., Song, F., Cameron, J., Stocum, D., Palakal, M. et al. (2010). "Systems biology approach to elucidate limb regeneration", Research Day, IUPUI.
- Jhamb, D., Rao, N., Milner, D., Li, B., Song, F., Wang, M., Voss, R., Palakal, M., King, M., Saranjami, B., Nye, H., Cameron, J., Stocum, D. et al. (2009). "Proteomic analysis of blastema formation in limb regeneration", Research Showcase, Institute for genomic biology, UIUC.
- Sanders, P., Jhamb, D., Palakal, M. (2009). "Computational analysis of biological networks", Research Day, IUPUI.
- Jhamb, D., King, M. (2008). "Proteomic analysis of changes during the onset of amphibian limb regeneration", Research Showcase, ISU.

Awards

- Indiana University Graduate School 2015 IUPUI Chancellor's Scholar
- Best Presentation Award – CAMDA, ISMB, 2013
- Best Poster Award - 2008, 2009, 2010
- Best Lecturer Award - Lecture Competition, Manipal University, 2005
- Best Student Award - Manipal University, 2004

Extracurricular Activities

- Member, Women in Technology, IUPUI, 2009 – Present
- Executive Member, Indian Student Advisory Council, IUPUI, 2009-2010