DATA MART INTEGRATION OF THE PROTEOME

Jay Vyas, Ph.d.

University of Connecticut, 2012

A broad range of tasks in modern bioinformatics analysis require integration of data from disparate sources. The explosion of data in the post-genomic era blazes a trail that for integrative bioinformatics: the use of disparate information repositories to solve problems in data visualization, interpretation, and normalization which have previously been difficult to address.

In order to integrate such repositories, we must maintain a dynamic data-integration framework that is capable of processing large amounts of data in an optimal manner. Although these requirements may be opposed, we can reconcile them by combining the attributes of a federated database environment with data marts: high-performance, task-specific databases which can be rapidly generated and torn down, due to their small footprint.

This thesis reveals the power of data marts for solving emergent problems in protein bioinformatics over a broad range, including functional annotation, the use of integrated methods for data visualization and interpretation of biomolecular data, and protein sequence mining. The broad range of examples demonstrate that data mart integration of the proteome is an efficient and practical alternative to monolithic approaches for integration.

DATA-MART INTEGRATION OF THE PROTEOME

Jay Vyas

B.S., University of Arizona, 2004

M.S., Rennsalear Polytechnic Institute, 2007

A Dissertation

Submitted in Partial Fulfillment of the

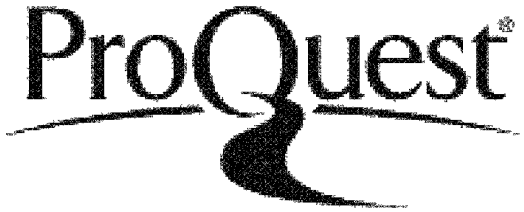
Requirements for the Degree of

Doctor of Philosophy

at the

University of Connecticut

2012



ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

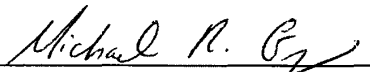
APPROVAL PAGE

Doctor of Philosophy Dissertation

DATA-MART INTEGRATION OF THE PROTEOME

Presented by

Jay Vyas, B.S., M.S.

Major Advisor 
Michael R. Gryk

Associate Advisor 
Mark Maciejewski

Associate Advisor 
Asis Das

University of Connecticut

2012

Acknowledgments

I would like to express my acknowledgements to the laboratory of Dr. Gryk for infinte patience, gentle guidance, and unwavering belief in the power of data integration.

Additionally I want to thank Dr. Schiller for sharing with me his encyclopedic knowledge of the proteome both in tabular as well as oral format,

Dr. Setlow for all the things that everybody normally thanks Peter for,

Dr. Hoch for his lending me use of his creative spirit and awe-inspring computional facilities,

Dr. King and Dr. Das for their proactive mentorship in the areas of molecular biology and biomolecular analysis,

Anne Cowan and John Carson for keeping an open mind at all times,

Mark Maciejewski for teaching me about shell scripts and CYANA, and **the late Steve Pfieffer – for making biology fun.**

Also I should thank all our collaborative departments, including the CCAM community, the Computer Science departments at Rennsalaer at Hartford and at Storrs, and all members of the Schiller Laboratory at UNLV.

Table of Contents

List of Tables iii
List of Figures iv
Introduction 1
Sequence Oriented Bioinformatics: The Early Years 1
From Bioinformatics to "Systems Biology" 3
Systems Biology and Protein Bioinformatics 5
The Current State of Protein Bioinformatics 6
The CONNJUR and MNM Projects Aim to Integrate Protein Bioinformatics 7
Scope of Study 9
Federated Systems and Data-marts: A Strategy for Data Integration 13
A Proposed Syntax for Minimotifs, Version 1 18
Mimosa: A Minimotif System for Annotation 31
VENN, a tool for titrating sequence conservation onto protein structure 41
Extremely Variable Conservation of y-Type Small, Acid-Soluble Proteins from Spores of Some Species in the Bacterial Order Bacillales 47
Chapter 5: The R3 Methodology for NMR Structure Calculation in Sparse Data 78
Appendix A. Additional Material for Chapter 1 101
Appendix B: Additional Material for Chapter 3 110

List of Tables

Table 1 Attributes of a minimotif definition 21

Table 2 Definitions of minimotif elements 22

Table 3 Residue frequencies in SH3 domain ligands 27

Table 4 Paper tracking status definitions 36

Table 5 Larger training set sizes (negative, positive) modestly improve algorithm performance 38

Table 6 CA-accuracy and heavy atom precision (reported by CYANA) for 600 structure calculation experiements 66

List of Figures

Fig. 1. An UML-attributed (www.uml.org) diagram exemplifying the fractionated yet interrelated nature of bioinformatics data repositories 17

Fig. 2. Entity-relationship diagram of a conceptual minimotif data model 23

Fig. 3. A physical implementation of the conceptual minimotif data model in MySQL. Relationships between tables are indicated 24

Fig. 4. SH3 binding minimotif family 26

Fig. 5. General architecture of MimoSA. 33

Fig. 6. Screen shots of MimoSA application database management windows ... 35

Fig. 7. Screenshot of MimoSA abstract and protein sequence viewers 36

Fig. 8. Screenshot of MimoSA paper browser and paper tracking windows 37

Fig. 9. ROC curve analysis of TextMine results 39

Fig. 10. Data processing model for VENN. 42

Fig. 11. Homology titration of C/EBPβ using VENN 43

Fig. 12. Comparison of ConSurf and Evolutionary Trace analysis of C/EBPβ 44

Fig. 13. Comparison of amino acid sequences of g-type SASP from spore-forming *Bacillales* 48

Fig. 14. Comparison of amino acid sequences of ab-type SASP from *A. acidocaldarius*, *B. subtilis*, *B. tusciae*, *G. kaustophilus* and *Paenibacillus* species 49

Fig. 15. Polyacrylamide gel electrophoresis at low pH of acetic acid extracts from spores of *P. polymyxa* ATCC 842 (lanes 1,2) and *A. acidocaldarius* NRS 1662. (lane 3) 50

Fig. 16. Putative upstream and downstream regulatory regions for genes encoding ab-type and g-type SASP from various species 51

Fig. 17. Phylogenetic tree for *Firmicute* species 52

Fig. 18. A comparison of control and R3 calculation accuracy when varying the amount of data for calculation (either chemical shifts or peaks) 65

Fig. 19. Control calculations vs rescue calculations visualized 72

Introduction

"The curtain was rising on the greatest show on earth."

-Russell Doolittle, From "The Roots of Bioinformatics In Protein Evolution" (2010).

Sequence Oriented Bioinformatics: The Early Years

Our ability to understand life on a molecular level would be impossible in the absence of key insights of the 20th century that correlated nucleotide sequences, protein sequences, protein structures, protein function, and cellular phenotype.

Scientists of the 1940s and 1950s were aggressively in pursuit of the relationship between DNA and heredity. These decades witnessed emergence of evidence linking DNA to the transformation of organism phenotype (McCarty and Avery 1946). While this implied that DNA was related to cellular function, it was not until the 1950s that clear evidence emerged for DNA's regular and linear structure (Watson and Crick 1953). These were seminal advances in the history of molecular biology (Lederberg 1994, Knight 1997). It was at this time, historically, that DNA took center stage in the quest to understand the molecular basis for cellular biology.

Parallel advances were occuring in the protein world. The importance of protein structures was demonstrated by X-ray methods in the middle 20th century demonstrating that proteins had a three-dimensional structure related to their

function (Muirhead and Perutz 1963). Scientists soon converged on the missing piece to the puzzle of how DNA, amino acids, and molecular function were interwoven with the discovery that simple nucleotide sequences are translated via a "genetic-code"(Nirenberg et al. 1965). There was now a theoretical basis for connecting genes, proteins, and cellular biology at a molecular level.

In 1973, these advances were complemented by work indicating that the amino acid sequence of a protein determined both its structure and subsequently, its function (Anfinsen 1973). It was now clear that genes, nucleotides, proteins, and their structures (as well as functions) were directly related. As computational technology continued improving during this time, scientists increasingly began utilizing linear sequences to characterize the molecular basis for life, and the field of bioinformatics was thus born.

Biological databases of the 1970s and 1980s enabled "sequence" mining as a new technique for knowledge acquisition (Dayhoff 1965, Doolittle 1981). These tools were now generating fascinating inferences on a regular basis. Emblamatic of this paradigm was the famous computational discovery of the link between cancer-causing agents in monkeys and human "growth factors" (Doolittle et al. 1983).

The usefulness of these tools was, although unquestionable, severely limited by the sparseness of high quality sequence data. The hundreds of sequences spanned by databases of this era represented but a fraction of the mammalian genome (which is comprised of tens of thousands of genes) (Doolittle 1981, Pruitt et al. 2005). The time-consuming and laborious nature of

gene and protein sequence data collection was a bottleneck to the growth of these repositories, and the curation and integration of such information from the literature represented another key struggle at the time (Doolittle 2010, Strasser 2010).

The Human Genome Project (HGP) was advocated by Jim Watson and others as a means to address the need for higher throughput, comprehensive accumulation of sequenced data. The HGP brought automated methods for sequence acquisition to bear - directly addressing the data-collection bottleneck of the 1980s, revealing the vast majority of the protein-coding content of the human genome (Ventner et al. 2001). This enabled and inspired analytical treatment of cellular systems on a much larger scale:

In this landmark study, Craig Venter's call to action resounds even today: *"All genes and their control elements must be identified; their functions, in concert as well as in isolation, defined; their sequence variation worldwide described; and the relation between genome variation and specific phenotypic characteristics determined. Now we know what we have to explain (Venter et al. 2001)."*

From Bioinformatics to "Systems Biology"

Ventner's words challenge us to understand and define the information contents of molecules in our cells on a genomic scale. This field of study is known as molecular "systems biology" (Kitano 2002, Peri 2003). The progression from molecular genomics to systems biology is natural: the biological molecules encoded by our genomes are now revealed; and we seek to

understand their function. But function is often relatively defined in biology, because molecules of the genome have coevolved specifically to functionally complement many intramolecular partners in the context of the cellular environment. Thus to characterize any one molecular component of the genome we must characterize the genomic and cellular systems in which that component operates.

Needless to say, the goals of systems biology are lofty. Genes are often defined in terms of their sequence, structural, and functional properties, and there are typically tens of thousands of genes encoded in any one mammalian genome. The integration of these attributes is essential to our understanding of biomolecular function (Kitano 2002, Boeckman 2005). We must thus interrelate semantically diverse biomolecular information on a grand scale if we are ever to precisely understand the role the thousands of molecules in encoded in our genome.

One could claim that such tasks are inherently digital (i.e. they are ideally suited suited by computers). First, a mammalian proteome alone comprises over 2 GB of plain text data, and if one were to include data records describing sequence, structural, and functional attributes, the amount of information rapidly scales. Second, we often aim to detect patterns in large, complex genomic data sets, and this requires that such data be structured in a way that renders it efficiently accessible to sophisticated computational methods – and of course computers are only capable of processing digital data. Third, biological knowledge is constantly evolving, and thus any attempt at capturing it must be

allow for rapid scanning and updating of information. The size, complexity, and dynamic nature of the genome's information content demand the use of computational methods for data management, acquisition, and analysis of genomic information.

Thus, the goal of identifying and defining the components of our genome depends on the availability of computational tools that integrate different types of biomolecular information on a large-scale. The integration and analysis of diverse moieties of biomolecular data remains an important paradigm in modern bioinformatics (Blundel et al. 2006, Stein et al. 2003, Venkatesh et al. 2002).

Systems Biology and Protein Bioinformatics

In order to provision the computational tools necessary to support the integrated analysis of biomolecular information, it is accepted that we should support the integration of the numerous categories and classes of biomolecular data processing tools and data types (Aasland 2002, Fox-Erlich et. al 2004, Gryk et. al 2010, Marchler-Bauer et al. 2003).

The majority of known functional genes in the genome are ultimately realized as proteins: if the genome contains the blueprints for our cells, then proteins are the actual buildings. The fundamental challenges in understanding systems biology are directly related to our comprehension of the roles that proteins play in the cell (Blundell et al. 2006, Boeckman et al. 2005, Kitano et al. 2002, Stark et al. 2005). It is obvious to say that this family of molecules is thus essential to our systematic understanding of life.

To this end, the research community has spent several decades on the digital curation, classification, and integration of protein data. Many proteins can now be classified and subdivided into representative structural and functional groups (Eckland et al. 2005, Marchler-Bauer et al. 2003). For example, by categorization of proteome's motif and domain elements, we now know that there exist approximately 500 different kinases, with a similar number of SH3 domains, encoded in our genomes. Such inferences are made possible by the continued expansion and curation of large, public bioinformatics databases (Manning 2002, Pruitt et al. 2005).

Proteins can be essentially defined in terms of their sequences. This definition is both efficient as well as useful –sequence begets protein structure that ultimately determines protein functionality. This functionality ultimately drives the cellular processes that we seek to understand (Anfinsen, 1973). The connection between protein sequences, protein structures, and molecular functions is now a critical "dogma" in biology.

The Current State of Protein Bioinformatics

Many existing pieces of work indicate that the integrative analysis of sequences, structures, and functions comprises a powerful technique for extracting knowledge regarding precise aspects of protein evolution and functional inference (Landau et al. 2005, Morgan et al. 2006). Unfortunately, this integration is not reflected in the way bioinformatics data for these records are managed (Blundle et al. 2006, Goble and Stevens 2008, Stein 2003). In order to carry out database driven protein analyses of this integrative nature, scientists

are forced to compensate for the disparate nature of bioinformatics data repositories (Stevens 2001, Saergent et al. 2011).

It is known that data integration is error-prone and time consuming in the biological sciences, especially when we consider data sets of genomic scale. The scenario of fractionated data renders the science of protein data integration a domain of science that many have sought to advance in recent years. The need for more comprehensive support of bioinformatics data integration of the proteome is currently acknowledged as a major issue (Reeves 2009). Some have attempted the streamlining of such tasks by brute force methods, which aim at creating data warehouses that integrate all data using a single information model – but such efforts were demonstrated to be fragile and unmaintainable (Stein 2003). Nevertheless, our ability to understand the key aspects of protein function hinges on our capacity to combine information from databases in a meaningful way.

The CONNJUR and MNM Projects Aim to Integrate Protein Bioinformatics

Two independent research initiatives were recently undertaken to improve the integrated analysis of protein data by providing support for integration: The CONNJUR (Connecticut Joint NMR University Research) and MNM (Minimotif Miner) projects. These are the founding projects behind this work, respectively aimed at facilitating a better description of proteins from the structural and functional standpoints (Gryk et al. 2010, Rajesekaran et al. 2009). In general, the projects both aim at modeling information in a precise manner as well as

provisioning tools that are readily applied to solving real world problems in sequence and structural bioinformatics.

The CONNJUR project focuses on integrating the process of protein structural analysis using NMR (Nuclear Magnetic Resonance), whereas the MNM project aims to catalog and facilitate the analysis of short, functional peptide segments of less than 13 amino acids (known as Minimotifs) which are conserved in eukaryotes which are (defined thoroughly in Chapter 1), for elucidation of modular, conserved functional subunits in full-length proteins. This thesis represents a fusion of concepts from these overall projects – applying the principals of the CONNJUR mandate for broad-scale integration of protein structural analysis workflows with MNM's goals of improvement of our ability to predict and define protein function in a broader biological context.

In this work, I have focused on a variety of emergent problems in the area of protein bioinformatics that are relevant to the above-mentioned projects. These include (a) the curation of protein functional annotations, (b) the visual interpretation of protein structures in an evolutionary context (to determine specificity and functional roles of molecules), (c) interactively locating and predicting the evolutionary origin of poorly conserved proteins, and (d) streamlining the NMR data processing workflow for structure calculation.

These tasks share a common attribute: the need for explicit, structured integration of protein data artifacts of varying types. *The thesis of this work is that explicit computational modelling and integration of protein data solves several emergent problems in protein bioinformatics, including the improvement*

of methods for Minimotif data curation, structural-functional analysis of proteins, protein derived NMR data processing, and inference of gene emergence. These problems are representative of a broader range of problems in bioinformatics which may be addressed in a similar manner.

Such advances in the computational treatment of these data types represent key steps in increasing our ability to extract knowledge from protein data archives. The models, principles, and strategies discussed in these pages thus intended to enable the protein bioinformatics infrastructures of the future, particularly those utilized in the CONNJUR and MNM initiatives.

Scope of Study

We aimed to develop practical solutions to emergent problems in protein bioinformatics in this work, specifically in the context of the CONNJUR and MNM projects. First, we develop and implement database integration frameworks which can be generally applied to the curation of Minimotifs in Chapters 1 and 2, wherein data from several sources is integrated in the service of sequence motif data management. Broadly applicable data-marts are designed using similar strategies in chapters 3 and 4. We apply the integrative techniques to the area of protein NMR in Chapter 5 so as to demonstrate a novel method for structure calculation. We briefly sum these chapters here:

In Chapter 1: "A Proposed Syntax for Minimotifs, Version 1", we encountered a need to integrate functional annotations for thousands of proteins for the Minimotif Miner database and application. Minimotifs are short, functional peptide segments that occur at high frequency in eukaryotic organisms, playing

many important roles in molecular interaction networks and other systems. The construction of a robust generic database of such Minimotifs is implemented using a precise, newly derived syntax for peptide function, via a database that is populated by a system for ingesting sequence, taxonomical, and literature-derived data. The value and robustness of this model for protein functional data is then demonstrated using an array of statistical analyses characterizing SH3 domains.

Chapter 2: "MIMOSA – A System for Minimotif Annotation" delineates the MIMOSA application for end-to-end curation of short, functional Minimotifs. The chapter builds on the work in Chapter 1. MIMOSA is optimized for end-to-end curation of thousands of functional peptides, or "Minimotifs," into an MNM database that will be heavily utilized. Through integration of Refseq, Pubmed, and several other algorithms and data sources, the MIMOSA application demonstrates a method of database integration for curation of large data sets.

In particular, MIMOSA extends the work in Chapter 1 by distilling the core database characteristics into a system that is focused on curating new Minimotifs found in the literature. Additionally, the MIMOSA system presents a novel algorithm for scoring text abstracts with respect to semantic content that is directly integrated into the curation system. By distilling the core components of our Minimotif functional model and automating curation, MIMOSA represents the first automated, end-to-end database for ingestion and processing of structured protein functional data, further validating the feasibility of the federation strategy and extending its scope to the domain of data warehousing and curation. This

chapter demonstrates a concrete application of the syntactical model for molecular function of Chapter 1, using a data integration approach that is flexible enough to support ongoing importation of novel data records over time.

Chapter 3: "Venn – A Tool for Titrating Sequence Conservation onto Protein Structures" presents the problem of integrating data of fundamentally different types (sequences, structures, and functions). This project represents important aspects of both the CONNJUR and MNM initiaitives, which aim to deal with many moieties of protein data in an integrated manner. In order to fuse such data, Venn heavily relies on database integration methods that are specific to a precise, data-driven workflow. This strategy was designed to enable real-time integration of protein structures with up-to-date sequence records available via web services. The particular workflow that Venn automates is now known as "homology titration." This method was utilized to reveal key specificity determinants in DNA binding which less robustly integrated analysis workflows are not capable of recovering. Iterative analysis enabled by higher levels of data integration is a fascinating paradigm in computational science that has many applications in bioinformatics, and is again visited in Chapter 5.

Venn's homology titration workflow would be highly impractical without an integrated methodology that supports automated ingestion of external proteomics repositories (i.e., EBI and the PDB) along with an internal representation of protein structural/sequence data – the PDB currently has tens of thousands of structures. Because of its federated nature, Venn was efficiently capable of running on a variety of platforms with an extremely small footprint.

In Chapter 4, "The Extremely Variable Conservation of γ-Type Small, Acid-Soluble Proteins from Spores of Some Species in the Bacterial Order Bacillales", the integration of phylogenetic data with interactive protein sequence scanning methods was utilized to bound the point of emergence of the variably conserved SSPE gene in gram-positive, sporulating bacteria. To survey the entire sequence space of such proteomes, a database was designed to interactively and comprehensively compare sequences in a controlled fashion, allowing for precise and interactive thresholding of sequence homology scans. Identifying all SSPEs in a subset of approximately 50 Firmicute proteomes and integrating the results with 16S RNA databases from other repositories enabled prediction of the point of emergence of the SSPE gene. Results were confirmed for these speculations using empirical techniques for protein identification in various species.

Chapter 5, "The R3 Methodology for NMR Structure Calculation in Sparse Data Backgrounds" describes an experimental addition to the CONNJUR framework for NMR data processing that enables calculation of protein structures on heavily pruned input data sets (that is, data sets where copious amounts of assigned chemical shifts and available NOESY peaks have been removed). This method was tested using an in-memory model of the structure calculation process that is capable of auto generating hundreds of test data sets as inputs to the traditional, semi-automated structure calculation process – which typically fails in sparse backgrounds. The large-scale automated testing of this method using the CONNJUR integration framework demonstrated the theoretical viability

of this new method for structure calculations, and is an important step forward towards the construction of a fully integrated solution to NMR structure calculation.

The approaches taken in these chapters demonstrate the key aspects of integrating the resources diagrammed in fig. 1.

Federated Systems and Data-marts: A Strategy for Data Integration

In this section, the aspects of database integration that are foundational to this research are discussed. All databases share a common thread: they provide access to some corpus of information in an organized and structured manner (Bergeron 2002, Simsion and Witt 2005). For example, a phone book might be thought of as a primitive database: It is used to collect and index a large body of information describing the locations of businesses and/or people.

Many digital bioinformatics databases exist today, cataloging a broad range of data about biological entities (Berman 2000, Pruitt et al. 2005, Sayers et al. 2010). Historically, such repositories have been "file" based: they accumulate records in large files, or clusters of files, and impose a higher order of organization on such files using folders, internal formats, or indices (Berman 2000, Pruitt 2005, Vyas 2008).

Modern structured databases (in particular, the "relational database") go one step further by storing data records as a decomposition of uniform, semantically meaningful relationships and attributes (Erlich et al. 2004). This strategy enables abstract operations on different records, which allow for

normalization and integration not easily achieved using a simple file-based approach.

Traditional relational databases may therefore play a key role in approaching more robust data integration by automating rich queries spanning different data types, but they are not, in and of themselves, a one-stop panacea for all data integration problems (Bergeron 2002, Simian and Witt 2005, Venkatesh et al. 2002). For example, relational systems do not natively support the access of data from fractionated and non-relationally structured sources. The distributed nature of modern bioinformatics databases thus requires a higher level for integration that goes beyond the decomposition and reformatting of information. This fractionated landscape is informally depicted in fig. 1.

An approach to integration of such fractionated resources commonly exists in one of two common forms: a "federated system" or a "data warehouse" (Bergeron 2002, Venkatesh et al 2002). A federated system is capable of serving data from a wide variety of sources via a simplified, central portal that links to external resources. Such a repository might be referred to as a "façade" or "proxy". Meanwhile, the data warehouse focuses on hosting such data by collocating it (Venkatesh et al. 2002). Thus, a warehouse directly integrates records (in contrast to the federated system, which "proxies" them). As one might expect, federated systems are ideal for synchronous access of rapidly changing data, whereas data warehouses excel in offline, analytical tasks.

A third construct for data integration, which is most representative of the approaches taken in this work, is known as the "data-mart". Data-marts are

small, efficient data warehouses that excel at a specific and well-defined task. Data-marts typically make up for their lack in completeness by realizing highly efficient, low-cost solutions to data mining problems that can be effectively bounded in scope (Bergeron 2002).

For the protein data integration tasks described herein, we heavily rely on federation to generate data-marts. Our reliance on federated methods is partially due to the overbearing constraints that data warehouses may impose in certain scenarios (Simsion and Witt 2005). For example, consider the task of creating a database of mammalian proteome records using a data warehouse: Such a construct, storing the entire sequenced proteomes of life, would contain upwards of approximately 2GB of raw textual sequence data and would consist of approximately 4 billion amino acids, only 5% of which would be mammalian (ftp://ftp.ncbi.nih.gov/refseq/release/release-statistics/). That is, 95% of the sequence portion of this database would be completely unused. It is obvious that (when dealing with data of this magnitude) analysis on a conventional computer might be suboptimal if a data warehouse strategy were to be blindly chosen in all cases where integration was required.

The need to understand the role of all genes, their interrelationships, and their particular functional attributes is lofty enough as is. There is obvious immediate value in further constraining scope of bioinformatics to a particular methodology or technical dogma. Thus, although many theoretical principles of database design are applied throughout this work, we focus more on analytical

support of the biomolecular data integration workflows, rather then the blind application of any one particular computational technique.

In this spirit, the following pages exemplify an agile combination of relational databases (for querying), data-marts (for integration) and federated systems (for "lazy" data ingestion), all applied to a broad range of problems in the protein bioinformatics regime.

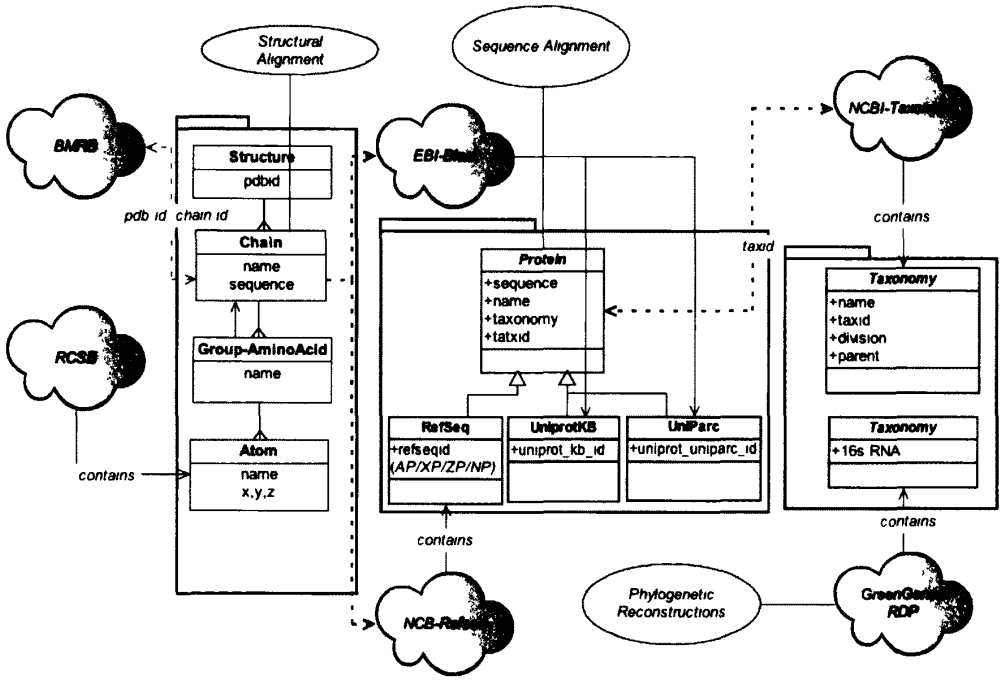


Fig. 1. An UML-attributed (www.uml.org) diagram exemplifying the fractionated yet interrelated nature of bioinformatics data repositories. Each square object (for example, "Protein") represents a data "class", which may have sub-types (i.e. UniprotKB), which comes from a data source (i.e. Refseq). Clouds represent data sources, dashed arrows represent relationships between data sources and data types (boxes), and straight arrows represent data "outputs" of clouds. In UML, "packages" can be used to separate different data classes, and that is done here to separate sequence, structure, and taxonomy. Left: Structural data comes from the Protein Data Bank (PDB). Empirical evidence for structures is stored at the Biomagnetic Resonance Bank, which can be linked to PDB chain ids. Structures can be compared using alignment methods. Middle: It is instructive to note that all "Protein" records that come from Refseq, UniprotKB, and UniParc may have sequences, names, taxonomies, and taxonomical ids – yet their identifier fields are distinct. External services for sequence alignment can match records to one another, and similarly, can be used to integrate structural data with sequence information (Chapter 3). Right: Taxonomical data can be integrated with genomic and protein sequence records using taxonomy identifiers, and compared using phylogenetic reconstruction algorithms (Chapter 4).

A Proposed Syntax for Minimotifs, Version 1

Published in BMC Genomics, 2009 : The databases, models, and queries presented in this work were designed using data marts which were designed and implemented by Jay Vyas. Other contributors are cited in the publication.

BMC Genomics

BioMed Cen

Research article

Open Acc

A proposed syntax for Minimotif Semantics, version 1

Jay Vyas[1], Ronald J Nowling[1], Mark W Maciejewski[1], Sanguthevar Rajasekaran[2], Michael R Gryk*[1] and Martin R Schiller*[1,3]

Address: [1]Department of Molecular, Microbial, and Structural Biology, University of Connecticut Health Center, 263 Farmington Ave. Farmington, CT 06030-3305 USA, [2]Department of Computer Science and Engineering, University of Connecticut, 371 Fairfield Rd., Storrs, CT 06269-2155 USA and [3]University of Nevada, Las Vegas, School of Life Sciences, 4505 Maryland Pkwy., Las Vegas, NV 89154-4004 USA

Email: Jay Vyas - jayunit100@gmail.com; Ronald J Nowling - rnowling@gmail.com; Mark W Maciejewski - madcm@nemesis.uchc.edu; Sanguthevar Rajasekaran - rajasek@engr.uconn.edu; Michael R Gryk* - gryk@uchc.edu; Martin R Schiller* - martin.schiller@unlv.edu

* Corresponding authors

Published: 5 August 2009

BMC Genomics 2009, 10:360 doi:10.1186/1471-2164-10-360

Received: 29 January 2009
Accepted: 5 August 2009

This article is available from: http://www.biomedcentral.com/1471-2164/10/360

Abstract

Background: One of the most important developments in bioinformatics over the past few decades has been the observation that short linear peptide sequences (minimotifs) mediate many classes of cellular functions such as protein-protein interactions, molecular trafficking and post-translational modifications. As both the creators and curators of a database which catalogues minimotifs, Minimotif Miner, the authors have a unique perspective on the commonalities of the many functional roles of minimotifs. There is an obvious usefulness in standardizing functional annotations both in allowing for the facile exchange of data between various bioinformatics resources, as well as the internal clustering of sets of related data elements. With these two purposes in mind, the authors provide a proposed syntax for minimotif semantics primarily useful for functional annotation.

Results: Herein, we present a structured syntax of minimotifs and their functional annotation. A syntax-based model of minimotif function with established minimotif sequence definitions was implemented using a relational database management system (RDBMS). To assess the usefulness of our standardized semantics, a series of database queries and stored procedures were used to classify SH3 domain binding minimotifs into 10 groups spanning 700 unique binding sequences.

Conclusion: Our derived minimotif syntax is currently being used to normalize minimotif covalent chemistry and functional definitions within the MnM database. Analysis of SH3 binding minimotif data spanning many different studies within our database reveals unique attributes and frequencies which can be used to classify different types of binding minimotifs. Implementation of the syntax in the relational database enables the application of many different analysis protocols of minimotif data and is an important tool that will help to better understand specificity of minimotif-driven molecular interactions with proteins.

Background

Minimotifs (also called Short Linear Motifs [SLIMs]), are short peptide sequences which play important roles in many cellular functions [1-3]. Many minimotif databases such as Minimotif Miner (MnM), Eukaryotic Linear Motif (ELM), phospho.ELM, DOMINO, MEROPS, PepCyber

and HPRD have cataloged more than a thousand minimotif entries and are expected to have significant growth in the near future [1,4-10]. Each of these databases model functional minimotifs in some capacity, often using individualized annotation schemes useful for the subset of minimotif data being managed. As the amount of minimotif data continues to grow, there are several expected advantages to be gained from the use of a standardized syntax. A standardized syntax will facilitate exchange of data with different minimotif databases. Likewise, a standardized syntax will allow integration with other non-motif databases enabling researchers to examine the connection of minimotifs with new types of data (e.g. disease mutations, protein structures, cellular activities, etc.), providing new opportunities for data mining. A standardized syntax will also allow refinement of minimotif sequence definitions, reduce redundant data, and normalize future annotation efforts.

The authors have been the curators of the Minimotif Miner database for the past four years. In compiling and managing this large dataset, we have had a lengthy and detailed exposure to the functional annotations currently reported in the scientific literature. This unique perspective has afforded us the insight as to certain common features of the functional annotation of minimotifs. Here we propose a standardized definition for minimotifs that is currently being used within MnM and which can be broadly applied to all minimotifs including those in the aforementioned databases.

We have observed that all minimotif annotations are composed of two major categories, the covalent chemistry and the function of the peptide. The first component of a minimotif definition includes its sequence and modification information. Schemes for modeling the sequence of minimotifs are well established and have been adopted from previous work modeling protein domains[11,12]. The protein sequences of minimotif instances are sequence strings of amino acids represented using an alphabet of IUPAC single letter code amino acid abbreviations [13]. For example, the 'PKTPAK' sequence in Kalirin describes an instance or single occurrence of a minimotif. Higher level minimotif abstractions are often represented as consensus sequences or position specific scoring matrices (PSSMs). Consensus sequence definitions identify permissible positional degeneracy. ProPxK is an example of consensus definition that describes multiple instances for proteins that bind to the SH3 domain of Crk; 'x' indicates that any of the 20 amino acids are allowed at the indicated position. Degeneracy can also be indicated for groups of amino acids that have similar chemical properties represented by a set of Greek symbols [14]. Consensus sequences can be represented as regular expressions in PROSITE syntax [12]. Probability-based PSSMs, like consensus sequences, represent the degeneracy at each position, but have the advantage that the probability of an amino acid at each position is explicit. PSSM are commonly represented as LOGO plots [15,16].

The sequence definitions described above, by themselves, have been found to be insufficient to describe many minimotifs which require additional covalent chemical modification. A set of rules for indicating post-translational modifications was previously defined by the Seefeld Convention [14]. One such rule is to indicate a phosphorylated residue by a lower case 'p' preceding an amino acid (e.g. RSxpSxP indicates the second Ser is phosphorylated in this 14-3-3 binding minimotif [17]). In our experience there are two important limitations imposed by the Seefeld Convention. First, the forced distinction between lowercase and uppercase character sets puts undesirable constraints on the implementation hardware/software; likewise the use of Greek characters to indicate degeneracy of amino acids with similar physical properties in minimotif definitions can also be problematic due to machine-specific character encoding. Second, this minimotif syntax is not extensible to all of the approximately 500 known posttranslational modifications, several of which have established roles in minimotif function [14,18]. For example, myristoylated residues and cis-proline bonds can not be enumerated using the Seefeld Convention. In this paper, we describe a model that overcomes these limitations for minimotif sequence definitions.

The second component of minimotifs is their biological function(s), which have generally been free-form descriptions in minimotif databases with no set standard. To our knowledge this minimotif subdomain of knowledge has not yet been modeled, which limits the ability to integrate data from different databases and hence their global usefulness. There are several ontologies that address domains related to minimotifs. The Gene Ontology (GO) defines a vocabulary for molecular and cellular functions and the association of these functions with gene products. While this ontology provides a useful resource for functional activities, the GO database is not designed to describe minimotif functions, nor capture important common attributes that are specific to minimotifs [19]. For example, the bind function in GO does not indicate the residues involved in an interaction, nor if any of these residues require any post-translational modifications. Likewise, the Protein Ontology, PSI-MOD, and RefSeq databases help to define entities that can be used for modeling minimotifs but are not sufficient by themselves for this purpose [20,21].

We provide a standardized semantic and syntactic definition of minimotifs gleaned from the data contained within MnM 2, and have executed its implementation by

BMC Genomics 2009, 10:360 http://www.biomedcentral.com/1471-2164/10/360

refactoring approximately 5000 minimotif annotations within MnM. As an example of the utility of this model and syntax, we demonstrate the use of the new database in classifying SH3 binding minimotifs.

Results

Minimotif Function Elements

A disambiguated and extensible semantic basis for minimotif functionality was derived from a set of rules which characterizes the approximately 5000 minimotifs in the Minimotif Miner (MnM) database [1] without information loss. We have not created a formal grammar, but rather a set of rules that characterize minimotif descriptions. For any minimotif clause, the syntax is *Minimotif* (subject), *Activity* (verb), and *Target* (object) which can be derived from a set of rules. We define these three major elements as follows:

Minimotifs consist of sequence definitions and sources. The sequence definition can be an instance, a consensus sequence, or a PSSM; all three classes of minimotifs are commonly reported in the literature. Instances represent primary data, whereas consensus sequences and PSSMs are interpretations of the data. *Minimotifs* may require one or more post-translation modifications such as phosphorylation or proline isomerization. In each motif, these modifications can be described by one or more residue names, type(s) of modification, and position(s) in the *Minimotif* sequence. Another approach for modeling residue modifications could be the atomic model previously described [22]. A source is the protein or peptide that contains the minimotif sequence. For example, in ' [PKTPAK in Kalirin] [binds] [Crk]', 'PKTPAK' is a sequence definition and 'Kalirin' is the minimotif source [23]. Alternatively, PxxPxK is a consensus definition that describes a consensus sequence for multiple instances.

Targets are proteins, nucleic acids, carbohydrates, lipids, small molecules, elements, metals, drugs, or complexes. In the case of proteins and nucleic acids, *Targets* may be associated with sequence definitions. *Target* proteins may contain domains as defined by the Conserved Domain Database [24], belong to a hierarchical classification based on fold [25] or refer to determined structure elements [26]. In the above example of the PKTPAK minimotif, the *Target* 'Crk' can be expanded to be more specific '1st SH3 domain of Crk'; referring to the N-terminal of two SH3 domains in Crk.

Activities are the actions of minimotifs and all minimotif activities can be generally classified as binds, modifies or traffics. The 'Binds' *Activity* describes an interaction of a protein containing a minimotif with another molecule. The 'Modifies' *Activity* defines a chemical change to a minimotif sequence that can be further subcategorized into enzymatic activities such as phosphorylates, amidates, geranyl gernaylates, cleaves etc The 'Traffics' *Activity* describes minimotif sequences required for a protein to be shuttled between cell compartments or other specific locations within or outside of cells.

In a number of minimotifs, a *Minimotif* and *Activity* are known, but the *Target* has not yet been identified or it is not yet known if the interaction of the *Minimotif* with the *Target* is direct. This information is still useful, thus we utilize a 'Required' *Activity* category which indicates that a minimotif sequence is necessary for a molecular or cellular activity. For example, the PNAY minimotif in Crk is required for Abl kinase activation [27]. In this case, Abl kinase activation is a subcategory of 'Required'. As in this example, the *Target* is null for the 'Required' *Activity*.

Minimotif Syntax

In order to combine these major minimotif elements and the minimotif sequence definition into human-interpretable semantic sentences we have defined 22 different attributes of minimotifs (Table 1) and derived the set of syntax rules listed below. Our goal was to identify a minimal set of rules that combine minimotif elements in order to regenerate valid minimotif sentences for the ~5000 minimotifs in the Minimotif Miner database. Valid minimotif sentences are based on these syntax rules, and biological entity categories of innumerable size (i.e. protein domains, protein names, molecule names, etc.).

Syntax Rules

Format: Minimotif elements in quotes are variable and defined in Table 1. Additional definitions are shown in Table 2. Bold text does not change and italicized elements are optional. Each minimotif function conforms to one of four rules (binds, modified, traffics, required).

'Minimotif' = 'Minimotif Sequence' (*'Required Modification'*) in 'Peptide' OR 'Protein'

'Protein target' = '*Domain position*' '*domain*' *domain of* 'Protein'

'Target' = 'Molecule' OR 'Protein target'

'Required modification' = '*Amino acid*' '*Position*' residue is '*posttranslational modification*'

'Activity modification' = '*Amino acid*' '*Position*' residue is '*posttranslational modification*'

BIND RULE: 'Minimotif' binds 'Target'

MODIFICATION RULE: 'Minimotif' is modified by the 'enzyme activity' of the 'Protein target' (*'activity modification'*).

Table 1: Attributes of a minimotif definition

#	Attribute[1]	Valid values and description
1	Motif sequence type	(Consensus, instance, PSSM) type of sequence definition
2	Motif sequence	Any consensus, instance, or PSSM describing a minimotif protein sequence
3	Required modification	description of chemical change to minimotif sequence
4	Motif source name	The name of protein or peptide that contains the minimotif
5	Motif source accession number	Swiss-Prot, RefSeq accession numbers for protein sequences containing the minimotif
6	Motif start position	Integer start position of the minimotif in motif source accession number
7	Motif source type	(Peptide and/or protein) indicates whether minimotif was investigated as a peptide fragment or in a protein domain
8	Activity	(binds, modifies, requires, traffics) the action of the minimotif
9	Subactivity	A more detailed description
10	Activity modification	Description of activity that covalently changes a minimotif sequence
11	Target name	The name of the molecule that acts upon the minimotif
12	Target accession number	If the target is a protein, the Swiss-Prot or RefSeq accession number(s) for Target protein sequence(s). The target can be a complex
13	Target type	(Peptide and/or protein) indicates whether *Target* was investigated as a peptide fragment or in a protein domain
14	Target domain	(any domain in the CDD) protein domain in the minimotif *Target*
15	Target domain position	Integer that indicates the relative location of a domain relative to its N-terminus for proteins that have more than one copy of the same domain
16	Target site	Integer for site where a minimotif binds a molecule, if more than one site is known
17	Subcellular localization	Region of the cell where the minimotif activity occurs
18	Affinity	(K_d, IC50, K_m) measurement of affinity of minimotif for its target
19	Structure	(PDB accession number) for a structure of the minimotif in complex with its target. A related attribute is 'related structures' of the minimotif source or target.
20	Experimental evidence	(X-ray, NMR, Phage display, peptide mapping, alanine scanning mutagenesis, evolutionary conservation, mutagenesis, modeling, deletion mapping, peptide binding, peptide competition, full-length protein, Surface Plasmon Resonance, ITC, SPOT array, Far-western, Co-Immunoprecipitation, yeast 2-hybrid, pulldown) different types of experimental evidence that supports a minimotif sentence.
21	Minimotif reference	(PubMed identifier or PDB accession number) indicates the references source(s) of the data supporting the minimotif definition
22	Database reference	Cross reference ID to other database that contains similar minimotif definition.

[1] Attributes are broken up into 4 sections related to the *Minimotif* (16), *Activity* (79), *Target* (1015), and properties (1619) of *Motif/Activity/Target* minimotif sentences.

BMC Genomics 2009, 10:360 http://www.biomedcentral.com/1471-2164/10/360

Table 2: Definitions of minimotif elements

Element	Definition
Minimotif	The covalent chemistry of a peptide segment represented by a sequence definition and any required modification and minimotif source
Minimotif sequence	An instance, consensus sequence, or PSSM that describes a peptide minimotif of less than 15 contiguous residues
Required modification	A change in the covalent chemistry of a minimotif sequence
Motif Source	The protein or peptide that contains the motif
Target	The molecule related to a minimotif by an activity
Activity	The action of the minimotif
Binds	Type of activity that involves a direct interaction between two or more molecule species
Modifies	Type of activity where the minimotif has a change in its covalent chemistry
Traffics	Type of activity where a protein moves between cellular compartments
Required	Type of activity where a minimotif is required for a chemical or cellular process
Chemical process	An event that results in a change of covalent bonds on a molecule
Cellular compartment	A place in the cell that can be discerned by the localization of at least one molecule
Peptide	Short polymer of amino acids
Protein	Polymer of amino acids
Domain	A region of a protein that folds independently.
Domain position	Location of a domain type in a protein that has more than one copy of a domain type relative to the N-terminus
Cellular process	An event or series of events that results in an observable change in a cell

TRAFFIC RULE: 'Minimotif' is trafficked by 'Target' to 'Cellular compartment' OR 'Minimotif' is trafficked to 'Cellular compartment'

REQUIRED RULE: 'Minimotif' is required for '*Chemical Process*' OR '*Cellular Process*'

Syntax Examples

BIND RULE: [IL]xxxxNPxY (tyrosine 497 residue is phosphorylated) in Interleukin 4 receptor binds PTB domain of IRS-1 [28].

MODIFICATION RULE: GRG in myelin basic protein is modified by the N arginine methylation activity of PRMT1 (Arginine 107 is methylated) [29].

TRAFFIC RULE: WHTL in Synaptotagmin is trafficked to synaptic vesicles [30].

REQUIRED RULE: GKFC in peptide is required for cell adhesion [31].

Minimotif Model and Implementation

The minimotif syntax was abstracted as a conceptual data model, which was used to derive logical and physical data models. An entity-relationship (ER) diagram of our conceptual data model is shown in Figure 1. The primary objects in the ER diagram are the *Minimotif* (green), *Activities* (orange), and *Target* (Cyan), each of which contains details regarding their attributes. Each *Minimotif* has a sequence and may have a modification (e.g. tyrosine phosphorylation in BIND RULE). All *Minimotifs* are in proteins which may have orthologues and domains. Each *Minimotif* can have a *Target* which is a molecule (Protein, Nucleic acid and small molecule are molecules; cyan). Molecules are in cell compartments. The *Target* has two relationships with the *Minimotif* (orange): modifies refers

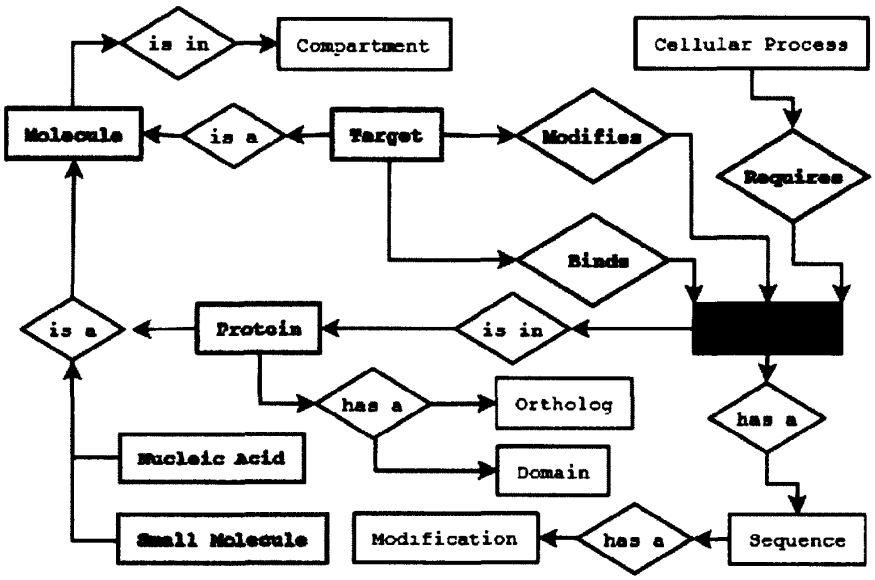


Figure 1
Entity-relationship diagram of a conceptual minimotif data model. Activities are colored orange; relationships are gray; molecules are cyan. There are properties of a *Motif/Activity/Target* in the database that are not present in this conceptual diagram

to a change in chemistry of the *Mnmmotif*, thus the *Target* is an enzyme in this case (MODIFIES RULE) For example, a *Mnmmotif* that is cut by a protease is chemically modified by an enzyme The *Target* can also bind the *Mnmmotif* (BIND RULE) In the case where a *Target* molecule is not known, the *Mnmmotif* may be required for some *Activity* as in the REQUIRED RULE above The TRAFFIC RULE is not represented in this diagram, but a *Mnmmotif* is trafficked by a *Target* from one cell compartment to another; the *Target* need not be known for the TRAFFIC RULE.

The physical implementation of the database is shown in Figure 2 The design of the mnmmotif relational database shows an intersection table (*motif_source*) of the *Mnmmotif*, *Activity*, and *Target* tables. Each mnmmotif in the database table has its own specific attributes such as mnmmotif type (consensus sequence or instance), a structure from the Protein Data Bank, an affinity for the *Mnmmotif/Target* complex, and published experimental techniques that support the *Mnmmotif/Activity/Target* relationship

We have previously reported the MnM 2 database which contains more than 5000 mnmmotifs [2] We have now refactored the MnM 2 database to use controlled vocabularies These include the Gene Ontology (GO, the *Activity* term names and id's for common molecular functions), NCBI Taxonomy for id's and species names, NCBI Conserved Domain Database (CDD, the names and identifiers for protein domains in motif *Targets*), NCBI Reference Sequences (RefSeq; for *Target* and *Mnmmotif* source protein names and ids), Human Proteome Organization (HUPO, for experimental evidence names and id's), PsiMod for post translational modifications of *Mnmmotifs*, and the Protein databank (PDB, for accession numbers for protein structure files) The new relational database that uses these controlled vocabularies enforces, normalizes, integrates, and explicitly defines the mnmmotif semantics. Details concerning the database are in Methods.

The mnmmotifs in the Mnmmotif Miner (MnM) database were refactored and implemented in MnM 2 [2] Our

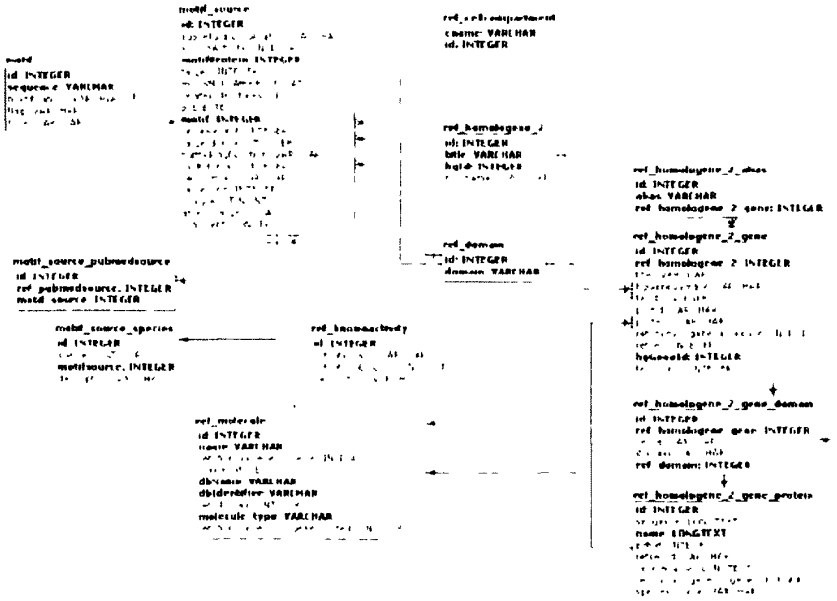


Figure 2
A physical implementation of the conceptual minimotif data model in MySQL. Relationships between tables are indicated. Three convergent lines pointing outward from a table indicate its dependency on another table. A circle or bar at the end of a line indicates that a relationship is optional or mandatory, respectively.

implementation of this model supports an integrative, semantically-rich minimotif analysis via the Structured Query Language (SQL), and importantly, is compatible with external motif analysis algorithms. This implementation enables extraction of groups of *Minimotifs* which share common values for any subset or combinations of subsets for the 22 different attributes in the model (Table 1) A set of 10 rules can be used to regenerate structured unambiguous human readable annotations (see Additional file 1)

We have built a user interface that enables users to query this database. This webpage is available as a link from the MnM 2 website. Users can select identifiers or text based descriptions from controlled vocabularies to query the database. For example, all SH3 binding motifs can by identified by selecting this domain from the CDD controlled vocabulary for domains [24] Many minimotif attributes can be queried from this page.

Once the query system is used to retrieve and group primary minimotif data (instances), interpretations of this data are often the next step in minimotif analysis. The interpretations of this data most commonly reported in the literature are consensus sequences, PSSMs, and groupings of families of minimotifs; these can be automatically generated based on query results generated by the aforementioned query system

Often a single laboratory does an experiment that identifies a consensus sequence, PSSM or grouping. MnM stores individual instances as reported in the literature, as well as inferred consensus sequences as reported by the authors. Our new query page has the advantage that consensus sequences, PSSMs or families of motifs can be generated from user-selected instances from one or more independent studies. Thus, this tool can be used to study groupings, consensus sequences, and PSSMs, which can vary significantly between different studies. Once groupings of

instances are selected from the new query page, users can then generate consensus sequences or PSSMs.

Grouping SH3 Domain Binding Minimotifs

There are many advantages expected to be gained by the use of a standardized minimotif syntax and query system. One such advantage is the simplified clustering of data within the database based on these new syntactical rules. As a case example, we classified 1363 SH3 binding minimotifs queried from the MnM 2 database. We selected this collection of data because of both the large number of reported SH3 binding minimotifs and the growing number of reported consensus sequences (e.g. PxxP, RxxPxxP, and PxxPxx [KR]). We posed a number of questions which would have been difficult to address without the syntax, but which are now easily addressed by querying the new relational database: Which SH3 consensus sequences are most common? How many SH3 binding consensuses are present in different instances? Do SH3 minimotifs bind to the same site? Is there a residue preference for degenerate positions?

A number of these questions had already been answered in an *ad hoc* fashion, but our goal in this case study was to address these questions in a systematic manner. Additional details for this analysis are provided [see Additional file 1].

The groups of SH3 binders were extracted by custom SQL statements filtering *Minimotifs* by type (consensus vs. instance), *Target* (SH3 containing proteins), and *Activity* (binds). This resulted in 1363 (741 unique) SH3 binding minimotifs, which could further be segregated into 69 consensus sequences and 672 instances. These sequences were compared inside our database for similarity based on the Shannon Information Content similarity metric as implemented by the Comparimotif library [32]. This analysis resulted in 10 minimotif groups that describe all SH3 binding minimotifs in the database (Figure 3). Details concerning the clustering analysis, queries, and results that lead to the distinct minimotif groups are provided [see Additional file 1].

Structural analysis of SH3 ligands

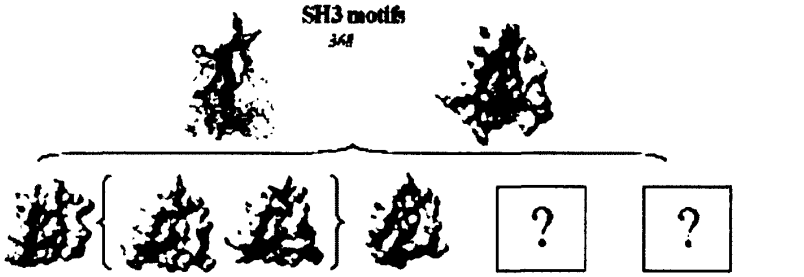
In order to better understand how these 10 SH3 binding minimotif groups were related to each other, we analyzed their known SH3/ligand complex structures. We queried the Minimotif Miner database and located representative structures for eight of the 10 groups. The *fit* function of Molmol was used to align the backbones of the eight SH3 domains using 6 residues in the β1 sheet, 4 residues in the 310 helix and 6 residues in the β4 sheet [33]. The root mean squared deviation (RMSD) for alignment of the backbone residues in these regions was 0.9 Å indicating a good alignment (Figure 2). We then examined the relationships of the binding sites of the different minimotifs by adding the sidechain bonds of the conserved residue positions and backbone atoms for each minimotif. For two structures we were only able to identify the binding sites based on nuclear magnetic resonance chemical shift mapping experiments [34,35].

Our analysis revealed that although SH3 domains are most commonly discussed for their ability to bind PxxP containing peptides, members of the SH3 domain family bind several different consensus sequences and have specialized structural interfaces. Of the 10 minimotif groups, many used different binding pockets on the SH3 domain. Four minimotifs bound in a similar region to the standard PxxP binding site (RxxPxxP, BxxB, PxxxPR, and KPTVY). The BxxB (B = basic) shares only one of two binding pockets with PxxP as previously noted [36,37]. Two of the motifs (RxxPxxP and PxxxPR) were found to bind in two different orientations with the peptides flipped ~180° in the binding sites. Two other consensus sequences bound previously identified alternative sites not near the PxxP site, and two had no structural information. This analysis confirms the distinction of the minimotif clusters derived by the sequence based-analysis.

Most SH3 domain binding peptides have multiple consensus sequences

Until recently, BxxB, PxxxPR, and several other types of SH3 binding minimotifs were not known. Given that there were 10 different types of SH3 binding consensus minimotifs, we wanted to know to what extent did previously studied ligands have multiple consensus sequences. We designed a query (query 9) that assessed how many consensus sequences were present in each ligand excluding the pairing of PxxP with RxxPxxP and PxxPx [KR] because these minimotifs are children of PxxP.

The average number of minimotif consensus per SH3 ligand was 2.3 indicating a tendency for each ligand sequence to have multiple SH3 consensus sequences. In the most extreme examples the SPTPPPVPRRGTHT, QPPVPSLPPRNIKP, KKPPPPVPKKPAKS, RRPPVPPR, and RRAPPPVPKKPAKG ligands each have five of the 10 different SH3 binding consensus sequences. For each consensus sequence, we have also reported the percent ambiguity in Figure 3 which is the percentage of each minimotif for which there are multiple consensus sequences. It is obvious from this analysis that a high proportion of previous SH3 binding experiments assessed ligands with potential to have multiple ligand binding modes. Thus, the majority of SH3 binding data may be subject to ambiguous interpretation (Figure 3). In interpreting many previous SH3 binding experiments, new ligand binding modes may now need to be considered in the experimental interpretation. Our database contains only 50 of the



?		PxxPxxP	PxxPx[KR]	PxxDY	KKPP
M	624	14.	343	3	3
A	52.5%	5.3%	6.5%	0.5%	0.5%

C	[illegible]	PxxxPR	KPTVY	RKxxYxxY	
M	91	21	2	3	1
A	29.9%	30.9%	11.5%	0.5%	0.5%

Figure 3
SH3 binding minimotif family. SH3 binding minimotifs were grouped into the 10 minimotif categories using the relational database and Shannon Information Content similarity metric. Surface plots of structures identified for 8 of the 10 group (1ZSG, black; 1NMF, pink; 1AZE, cyan; 2BZR, blue; 1CKA, magenta; 1OPL, red; 1BLO orange; 1HRH, green; 1NYG, brown) are shown. The carbon backbones of SH3 domains were fit using Pymol with residues in the β1 and β4 sheets, and the 310 helix, to an RMSD of 0.9 [33]. An overlay of each SH3 domain carbon backbone with its peptide minimotif is color matched and relevant minimotif side chain bonds are represented as thickened lines; the surface plot for the overlay is derived from the 1ZSG structure). Structures of the ligands for the RKxxYxxY and WxxxFxxLE minimotifs are not known, but the binding sites on the SH3 domains derived from NMR chemical shift mapping experiments are indicated. RxxPxxP and PxxxPR minimotifs show structures with the peptides in opposing orientations. The consensus sequences (C), total number of minimotifs for C (M), and percentage of potentially ambiguous ligand instances (A) in the MnM 2 database are indicated.

270 known human proteins with SH3 domains, thus the 10 SH3 minimotif groups we identified may become even more complex with a comprehensive analysis of all SH3 domains.

All SH3 domain binding peptides have basic residues

To further characterize the SH3 binding landscape, we performed analysis of residue content in all SH3 ligands using queries as described in methods. Compositional analysis showed a high preference for proline (4.2 fold), arginine (1.7 fold), and lysine (1.8 fold)(Table 3). In fact, all SH3 ligands in the database contained either a lysine or arginine, suggesting that a positive charge may be an important factor in ligand binding to SH3 domains. Another study has previously suggested a role for positively charged residues in SH3 domain interactions [38]. Consistent with this observation, the least enriched residues in SH3 ligands were the negatively charged residues.

The overall average calculated charge of SH3-binding peptides in our database was +3.2 ± 1.4 (average length of 12.1 ± 3.1 residues); this calculation is based on summing

charges of basic and acid residues assuming a neutral pH. Of nine other groups of minimotifs with common domain targets in MnM 2 only minimotifs for Calmodulin (n = 31) and 14-3-3 (n = 44) had net positive charges of 3.0 ± 1.3 and 1.0 ± 0.9, respectively; PDZ (n = 1089), SH2 (n = 952), kinase (n = 206), PTB (n = 168), protease (n = 93), FHA (n = 67), WW (n = 27) and phosphatase (n = 25) domains had ligands or substrates with an average neutral or net negative charge.

Collectively, these query results strongly suggest that known SH3 peptide ligands have a more positive overall charge than proteins in the human proteome. It is important to note that when restricting the SH3 ligand query to non-BxxB sequences, the average ligand charge was still +2.2 ± 1.2. Only 11 of the 1363 sequences had a neutral or negative charge and several of these were for WxxxPxxLE and PxxDY minimotifs, which have few instances in the dataset.

Discussion

We have developed a syntax with a set of rules that describes the more than 5000 minimotifs in the MnM database. While this syntax is complete for the data currently managed by MnM, we will actively continue to develop and expand this model to support additional types of data. The syntax is important because it enables the use of controlled vocabularies through defined rules, integration with other types of databases, exchange of data between minimotif databases, and the ability to address difficult questions that are facilitated through mining of minimotif data.

Current approaches for defining the covalent chemistry of minimotifs are not without limitations, beyond the post-translational modifications discussed earlier. The most commonly used representation of a motif is a consensus sequence. The definition of the word consensus does not necessitate that all members of a group conform, thus consensus sequences, while having the advantage that they can be used to group a number of instances, can also introduce ambiguity. For example, Calmodulin binding minimotifs have several members that do not conform to consensus sequences [39].

We have decided not to model a relationship between instances and their consensus sequences because these can be reconstructed through database queries that use a wider set of data. However, this approach remains to be tested with rigor and consensus sequences with nonconforming members may prove difficult. There are likely to be other ways that consensus sequences are limiting, for example, our SH3 minimotif analysis suggests that this binding minimotif should have an overall positive charge, which can not be represented by a consensus sequence. Furthermore, our semantics currently rely on consensus sequence definitions and our syntax does not support PSSMs. While a thorough discussion of sequence definition limitations is beyond the scope of this paper, we expect that through continued annotation using our standardized syntax we will able to identify all anomalies in our model and adjust it accordingly.

Through our work on minimotifs, we recognized a number of other important limitations that will need to

Table 3: Residue frequencies in SH3 domain ligands

Residue	Total Count	Composition (%)	Enrichment (fold)
A	554	7.4	1.0
C	118	1.6	0.7
D	102	1.4	0.3
E	100	1.3	0.2
F	206	2.8	0.8
G	275	3.7	0.6
H	54	0.7	0.3
I	171	2.3	0.6
K	764	10.2	1.0
L	697	9.3	1.0
M	64	0.9	0.4
N	150	2.0	0.6
P	2026	27.2	4.2
Q	200	2.7	0.6
R	752	10.1	1.7
S	404	5.4	0.7
T	310	4.2	0.8
V	336	4.5	0.8
W	59	0.8	0.7
Y	102	1.4	0.5

be addressed in the future. Several attributes of minimotifs could be modelled better. For example, some *Targets* of motifs are complexes, rather than single proteins. Furthermore, a specific structural conformation of a protein may be specific to a *Minimotif* or *Target*. Wherever possible we have tried to use controlled vocabularies, but a number of attributes could expand on this theme. We could better use vocabularies for activities and subcellular localizations from the GO database. However, we have recognized that all minimotif, and perhaps molecular activities, fit into the general categories of binds, modifies, or traffics, a basic grouping of function not implemented in GO. Alias names of proteins also present a problem with redundancies, but this is a problem endemic to many biological databases. While many previous minimotif descriptions in the literature use elements of the syntax we propose, the syntax is not always structured the same way, making automated annotation or restructuring of previous literature difficult. Finally, there is no guarantee that all future minimotif functions we identify will fit in our model.

We have shown that implementation of the syntax is useful. Our analysis of SH3 binding minimotifs identified over 1000 minimotifs that cluster into 10 major groups. The majority of these groups bound to a similar site but, the specific contacts in the interaction were generally not conserved between groups. Thus, it seems that while the evolutionary pressure for binding to the SH3 domain is strong, the precise mechanism of binding can vary. This SH3 minimotif analysis emphasizes the necessity of standardizing minimotif semantics and sequences in a well-modeled database with a query system that can be used to manage data from a collection of related studies. The data-driven classification provides a solution to grouping minimotifs based on a broad collection of experiments with reduced bias towards any individual peptide screen or study. The semantics and relational database are important in this process because a large amount of data can be normalized and because sequence similarity is not the only indicator of functional similarity. For example, PLPP and SKSKDRYY possess similar activities even though they do not share a single residue in common [40,41].

Conclusion

Information inconsistency arising from informal semantics is always a limitation for data integration. The minimotif semantics described here, along with the data model and its implementation, enable the computation of functional equivalence between minimotifs. This linguistic scheme is similar to one recently suggested by Gimona [42].

The syntax will facilitate many types of computational analyses of minimotifs. We are now able to generate specific subsets of data based on any of the 22 attributes of minimotifs. For example, the database facilitates refining sequence definitions similar to the recent refinement of a sumoylation minimotif [43]. The normalized syntax will allow exchange of data with other databases, reduce redundancies, and provides a framework for future annotations. The syntax also facilitates minimotif classification, as done for SH3 domain binding minimotifs in this paper.

Methods

Database Design

Our theoretical model of minimotif semantics is only useful if it is logically understood by a machine, thus the reason why we built a relational database. It is typical to implement database relationships in ways which exceed the complexity of the theoretical data model on which they are based (for performance and practicality reasons). Because many *Targets* can also be *Minimotif* containing proteins, and the three *Minimotif/Activity/Target* components are only related by experimental work, many additional tables were needed to link information for these components.

Full database documentation is provided [see Additional file 2]. Since the most important elements of our database are those which directly model the semantics, a mapping between our conceptual model and its physical implementation is provided in a table in Additional file 1. The physical model also includes many other federated data sources which are not in the conceptual model such as the gene alias names (ref_homologene_2_gene_alias), and minimotif annotation literature sources (motif_source_pubmedsource) which are linked to the ref_pubmedsource table (not shown). More information regarding these relationships is in Additional file 2.

Additional tables in the database were used for data mining. For example, Motif_source_motif_group groups minimotif_source records and ref_amino_acid is a table of all amino acids. The motif table contains the minimotif amino acid sequence and any post-translational modification to the sequence. Each minimotif is associated with one motif_source record, which is an intersection point for two ref_molecule records (one being the minimotif containing protein, and one being the molecule type of the target which the minimotif acts upon). The target is optional depending on the annotation rule.

Each ref_molecule entry can be optionally associated with either a RefSeq protein and/or a HomoloGene cluster, and additionally may have a ref_domain record (which is a federation of the NCBI Conserved Domain Database (CDD)) [24]. These clusters are important because many minimotif functions are conserved across species bound-

aries, allowing us to group RefSeq proteins which serve as minimotif targets.

Clustering of SH3 minimotifs [see Additional file 1]

Authors' contributions

JV, RJN, MRS, MRG, and SR developed the minimotif semantic and syntax. JV, MRS, MRG, and MWM contributed to the design of the minimotif data model. JV implemented the data model in a MySQL database. Refactoring and annotation of minimotifs into the minimotif data model was carried out by MRS. JV and MRS conducted the analysis of SH3 binding minimotifs. MRS, MRG, and JV prepared the manuscript and all authors were involved in editing. All authors read and approved the final version of the manuscript.

Additional material

Additional file 1
Supplementary Methods, Data, and Results. Supplementary methods and results for database design and clustering SH3 binding minimotifs.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-360-S1.doc]

Additional file 2
Database Documentation files. File of documentation of the MySQL data model.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-360-S2.zip]

Acknowledgements

We thank the National Institutes of Health for funding (GM079689). We would like to thank Dr. Stephen King and members of the Minimotif Miner team for suggestions in preparation of this manuscript.

References

1. Balla S, Thapar V, Luong T, Faghri T, Huang CH, Rajasekaran S, del Campo JJ, Shin JH, Mohler WA, Maciejewski MW, Gryk M, Piccirillo B, Schiller SR, Schiller MR: Minimotif Miner, a tool for investigating protein function. *Nat Methods* 2006, 3:175-177.
2. Rajasekaran S, Balla S, Gradie P, Gryk MR, Kadaveru K, Kundeti V, Maciejewski MW, Mi T, Rubino N, Vyas J, Schiller MR: Minimotif miner 2nd release: a database and web system for motif search. *Nuc Acids Res* 2009, 37:D185-D190.
3. Neduva V, Russell RB: DILIMOT: discovery of linear motifs in proteins. *Nucleic Acids Res* 2006, 34:W350-W355.
4. Diella F, Gould CM, Chica C, Via A, Gibson TJ: Phospho.ELM: a database of phosphorylation sites update 2008. *Nucleic Acids Res* 2008, 36:D240-D244.
5. Diella F, Cameron S, Gemund C, Linding R, Via A, Kuster B, Sicheritz-Ponten T, Blom N, Gibson TJ: Phospho.ELM: A database of experimentally verified phosphorylation sites in eukaryotic proteins. *Bmc Bioinformatics* 2004, 5:79.
6. Ceol A, Chatr-Aryamontri A, Santonico E, Sacco R, Castagnoli L, Cesareni G: DOMINO: a database of domain-peptide interactions. *Nucleic Acids Res* 2007, 35:D557-D560.
7. Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi TKB, Gronborg M, Ibarrola N, Deshpande N, Shanker K, Shivashankar HN, Rashmi BP, Ramya MA, Zhao ZX, Chandrika KN, Padma N, Harsha HC, Yatish AJ, Kavitha MP, Menezes M, Choudhury DR, Suresh S, Ghosh N, Saravana R, Chandran S, Krishna S, Joy M, Anand SK, Madavan V, Joseph A, Wong GW, Schiemann WP, Constantinescu SN, Huang LL, Khosravi-Far R, Steen H, Tewari M, Ghaffari S, Blobe GC, Dang CV, Garcia JGN, Pevsner J, Jensen ON, Roepstorff P, Deshpande KS, Chinnaiyan AM, Hamosh A, Chakravarti A, Pandey A: Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res* 2003, 13:2363-2371.
8. Gong WM, Zhou DH, Ren YL, Wang YJ, Zuo ZX, Shen YP, Xiao FF, Zhu Q, Hong AL, Zhou X, Gao XL, Li TB: PepCyber: P~PEP: a database of human protein-protein interactions mediated by phosphoprotein-binding domains. *Nucleic Acids Res* 2008, 36:D679-D683.
9. Puntervoll P, Linding R, Gemund C, Chabanis-Davidson S, Mattingsdal M, Cameron S, Martin DMA, Ausiello G, Brannetti B, Costantini A, Ferre F, Maselli V, Via A, Cesareni G, Diella F, Superti-Furga G, Wyrwicz L, Ramu C, McGuigan C, Gudavalli R, Letunic I, Bork P, Rychlewski L, Kuster B, Helmer-Citterich M, Hunter WN, Aasland R, Gibson TJ: ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res* 2003, 31:3625-3630.
10. Rawlings ND, Morton FR, Barrett AJ: MEROPS: the peptidase database. *Nucleic Acids Res* 2006, 34:D270-D272.
11. Gribskov M, McLachlan AD, Eisenberg D: Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci USA* 1987, 84:4355-4358.
12. Falquet L, Pagni M, Bucher P, Hulo N, Sigrist CJ, Hofmann K, Bairoch A: The PROSITE database, its status in 2002. *Nucleic Acids Res* 2002, 30:235-238.
13. IUPAC-IUB Commission on Biochemical Nomenclature (CBN). Abbreviations and symbols for nucleic acids, polynucleotides and their constituents. Recommendations 1970. *Biochem J* 1970, 120:449-454.
14. Aasland R, Abrams C, Ampe C, Ball LJ, Bedford MT, Cesareni G, Gimona M, Hurley JH, Jarchau T, Lehto VP, Lemmon MA, Linding R, Mayer BJ, Nagai M, Sudol M, Walter U, Winder SJ: Normalization of nomenclature for peptide motifs as ligands of modular protein domains. *FEBS Lett* 2002, 513:141-144.
15. Obenauer JC, Cantley LC, Yaffe MB: Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res* 2003, 31:3635-3641.
16. Schneider TD, Stephens RM: Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* 1990, 18:6097-6100.
17. Muslin AJ, Tanner JW, Allen PM, Shaw AS: Interaction of 14-3-3 with signaling proteins is mediated by the recognition of phosphoserine. *Cell* 1996, 84:889-897.
18. Garavelli JS: The RESID Database of protein structure modifications. *Nuc Acids Res* 1999, 27:198-199.
19. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: Gene Ontology: tool for the unification of biology. *Nat Genet* 2000, 25:25-29.
20. Wheeler DL, Chappey C, Lash AE, Leipe DD, Madden TL, Schuler GD, Tatusova TA, Rapp BA: Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2000, 28:10-14.
21. Sidhu A, Dillon TS, Sidhu BS, Chang E: Protein ontology: Seamless protein data integration. *Molecular & Cellular Proteomics* 2005, 4:684.
22. Fox-Erlich S, Martyn TO, Ellis HJC, Gryk MR: Delineation and analysis of the conceptual data model implied by the "IUPAC Recommendations for Biochemical Nomenclature". *Protein Science* 2004, 13:2559-2563.
23. Schiller MR, Chakrabarti K, King GF, Schiller NI, Eipper BA, Maciejewski MW: Regulation of RhoGEF activity by intramolecular and intermolecular-SH3 interactions. *J Biol Chem* 2006, 281:17774-17786.
24. Marchler-Bauer A, Anderson JB, DeWeese-Scott C, Fedorova ND, Geer LY, He S, Hurwitz DI, Jackson JD, Jacobs AR, Lanczycki CJ, Liebert CA, Liu C, Madej T, Marchler GH, Mazumder R, Nikolskaya AN, Panchenko AR, Rao BS, Shoemaker BA, Simonyan V, Song JS, Thiessen PA, Vasudevan S, Wang Y, Yamashita RA, Yin JJ, Bryant SH: CDD: a

Mimosa: A Minimotif System for Annotation

Published in BMC Bioinformatics, 2010

The application and software architecture presented in this work was architected and implemented by Jay Vyas. The TextMine algorithm was entirely designed and implemented by Jay Vyas.

Vyas et al. BMC Bioinformatics 2010, 11:328
http://www.biomedcentral.com/1471-2105/11/328

BMC Bioinformatics

SOFTWARE Open Access

MimoSA: a system for minimotif annotation

Jay Vyas[†1], Ronald J Nowling[†1], Thomas Meusburger[2], David Sargeant[2], Krishna Kadaveru[1], Michael R Gryk[1], Vamsi Kundeti[3], Sanguthevar Rajasekaran[3] and Martin R Schiller[*1,2]

Abstract

Background: Minimotifs are short peptide sequences within one protein, which are recognized by other proteins or molecules. While there are now several minimotif databases, they are incomplete. There are reports of many minimotifs in the primary literature, which have yet to be annotated, while entirely novel minimotifs continue to be published on a weekly basis. Our recently proposed function and sequence syntax for minimotifs enables us to build a general tool that will facilitate structured annotation and management of minimotif data from the biomedical literature.

Results: We have built the MimoSA application for minimotif annotation. The application supports management of the Minimotif Miner database, literature tracking, and annotation of new minimotifs. MimoSA enables the visualization, organization, selection and editing functions of minimotifs and their attributes in the MnM database. For the literature components, Mimosa provides paper status tracking and scoring of papers for annotation through a freely available machine learning approach, which is based on word correlation. The paper scoring algorithm is also available as a separate program, TextMine. Form-driven annotation of minimotif attributes enables entry of new minimotifs into the MnM database. Several supporting features increase the efficiency of annotation. The layered architecture of MimoSA allows for extensibility by separating the functions of paper scoring, minimotif visualization, and database management. MimoSA is readily adaptable to other annotation efforts that manually curate literature into a MySQL database.

Conclusions: MimoSA is an extensible application that facilitates minimotif annotation and integrates with the Minimotif Miner database. We have built MimoSA as an application that integrates dynamic abstract scoring with a high performance relational model of minimotif syntax. MimoSA's TextMine, an efficient paper-scoring algorithm, can be used to dynamically rank papers with respect to context.

Background

Minimotifs are short peptide sequences that are the recognition elements for many protein functions. These short sequences are responsible for protein interaction interfaces involving other proteins (or molecules) in cells, trafficking proteins to specific cellular compartments, or serving as the basis for enzymes to post-translationally modify the minimotif sequence. At present, many minimotif instances and consensus sequences are collected into a wide spanning set of relatively small databases such as MnM, ELM, Domino, PepCyber, and ScanSite [1-5]. Most databases focus on specific subsets of minimotifs. For example, Phospho-ELM has merged with PhosphoBase as a database that focuses on instances of phosphorylation on proteins [6]. Likewise, ScanSite largely concentrates on protein interaction minimotifs for a small subset of domains. In addition to these databases, recent years have seen increased publication rates of high throughput studies that generate minimotif data. Despite this growth in information, many of the reported minimotif attributes have yet to be integrated into any database.

The goal of the MnM project is to integrate well-structured data for a set of defined attributes of minimotifs in a single, non-redundant data repository with high accuracy. The number of reports of minimotifs in the literature has continued to grow since the late 1980's, recently with more rapid growth due to high throughput functional peptide screens. Previously, we showed that the several thousand minimotifs in MnM can be discretized into a structured syntax which can be directly enforced and modeled in a relational database [1,7]. Through this

* Correspondence: martin.schiller@uri.edu
[1] Department of Molecular, Microbial, and Structural Biology, University of Connecticut Health Center, 263 Farmington Ave. Farmington, CT 06030-3305 USA
† Contributed equally
Full list of author information is available at the end of the article

process, we recognized the need for a system that manages minimotif annotation, which would help identify papers, reduce the time required for manual annotation, reduce errors, duplications and ambiguities, and aids in maintenance of the database.

Currently, there are no bioinformatics tools designed for annotating minimotifs from the literature. Most reported annotation methodologies concentrate mainly on genomes and proteome scale data [8-10]. A proposed stratification of annotation efforts refers to sequence-based annotation as the first dimension of genome annotation which defines components [11]. The second dimension can be considered those annotations that focus on component interactions. This is exemplified by the human kinome and other types of functional annotations in the SwissProt and Entrez Gene databases [12,13]. Annotation of minimotifs can be considered a second dimension annotation.

In considering whether to design a novel minimotif annotation system or adapt an existing annotation system used for another purpose, we identified a number of requirements to facilitate accurate, non-redundant, and efficient annotation of minimotif literature. We wanted the system to interface with a relational database that enforces controlled vocabularies from external databases and eliminates duplication. The system should be able to read, write, and edit entries in a database. The system should display papers that have been and are yet to be annotated, as well as support database-driven machine learning that scores papers for minimotif content, paper sorting, and paper filtering. The system should also have the capability to track annotations from multiple annotators. Finally, the system should be capable of accepting the fine-grained information content of minimotifs, in a structured and comprehensive manner.

Despite advances in management and mining of scientific literature, no tool existed that met the requirements we required for accurately annotating minimotif data. For example, each of the existing annotation tools such as MIMAS, Textpresso and Biorat only addresses a subset of the above requirements [14-16].

In this paper, we describe MimoSA, a Minimotif System for Annotation designed for managing and facilitating minimotif annotation. MimoSA allows for minimotif-centric analysis of PubMed abstracts and annotation of minimotifs. MimoSA's contents are entirely database driven, thus enabling its adaption as an annotation tool for other information spaces that require extraction of information from the primary literature.

Implementation

We present the generalizable architecture and implementation of MimoSA, an application, which allows minimotif annotations to be entered, reviewed, edited, approved by multiple users, and disseminated through the publically-available MnM web application. We also describe a generalizable paper-scoring algorithm and its implementation for ranking papers that contain minimotifs. By embedding this methodology into MimoSA, PubMed abstracts can be scored and associated papers can be ranked based on the presence of minimotif information content.

MimoSA was developed in Java http://java.sun.com and interfaced with a MySQL database http://www.mysql.com using the Hibernate object-relational mapper http://www.hibernate.org. MimoSA was built to interface with the MnM relational database, which has been expanded to include the ability to store information about papers to be annotated and the relationships between minimotif annotations and their source papers [7]. The graphical user interface (GUI) was developed using Swing http://java.sun.com/docs/books/tutorial/uiswing. Supporting applications used for offline data processing were also developed in Java. These applications identify new keywords and terms used to highlight text in the abstract display window and download content and metadata from PubMed for papers added into the system. For these features, we have relied extensively on the PubMed Application Programming Interface (API) and Remote Procedure Call (RPC) library.

Unlike other annotation and text mining systems, the data artifacts produced by MimoSA are accessible by an API, which is syntax-driven and strongly typed. This allows for high-precision annotation of articles that is not coupled to any one data repository. Thus, MimoSA may easily be configured, for example, to save annotations to an XML document or text file by simply modifying the data access layer implementation.

The generality of the MimoSA application enables its adaptation to other databases and other knowledge domains. This was a consideration made during the development of MimoSA, so as to more broadly enable adaption to other bioinformatics projects.

Results

MimoSA prototype design

The primary function of MimoSA is to support the process of annotating functional minimotifs and their metadata from the primary literature. Secondary functions include minimizing user errors and data redundancy, improving annotation efficiency through techniques such as automated motif/activity/target suggestions, and aiding in the identification of papers containing minimotif content through a machine learning-based ranking system. MimoSA features distinct components and algorithms, which streamline these processes.

The general annotation workflow is as follows (see Fig. 1): Using the MimoSA client software, the annotator

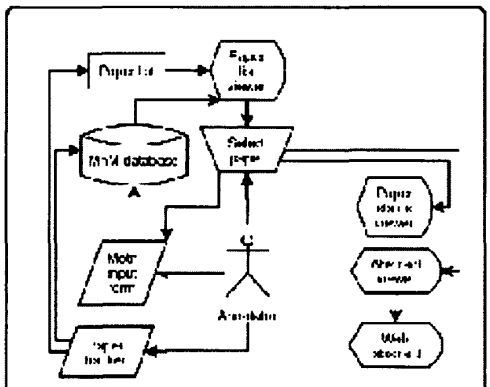


Figure 1 General architecture of MimoSA

accesses the server housing the MnM database. The user selects a paper for annotation using the Paper List Viewer. Selection of a paper automatically triggers the opening of the Abstract Viewer and the Minimotif Annotation Form and directs an external web browser to online versions of the abstract and full text paper, if available. Based on the information in the viewers, the Minimotif Annotation Form is used to modify an existing or enter a new minimotif annotation, which is then committed to the database. The annotation status of the paper is updated using the Paper Tracker Form.

The components of MimoSA can be broken up into three functional categories: MnM database management tools, minimotif annotation tools, and paper management tools. Descriptions of each component follow.

The database management tools consist of a minimotif browser and a minimotif editor. The minimotif browser shown in Fig. 2A displays all minimotif annotations in the MnM database and associated attributes in a scrollable window that also displays the total number of minimotifs. A Paper Browser is accessed from a tab and gives a list of papers that need annotation. From the paper or minimotif browsers, a Minimotif Annotation Form can be launched by double clicking a row to enter a new or modify an existing minimotif annotation (Fig. 2B-2D). This opens a tabbed frame where all the minimotif attributes are displayed and can be added or changed. Minimotif annotations can be selected for exportation as Comma-Separated Value (CSV) files for external manipulation. Likewise, an import function allows import from a CSV file. The minimotif annotations in the browser can be sorted based on a number of different attributes from a drop-down menu.

The minimotif annotation tools consist of the Minimotif Annotation Form, the Abstract Viewer, and the Protein Sequence Validator. Multiple forms can be displayed at once. On the Minimotif Annotation Form, there is a "clone" function, which opens a new instance of the form pre-filled with all of the minimotif-syntactical attributes except the minimotif's sequence and position. This is intended to facilitate more efficient annotation of high-throughput papers for minimotif discovery (e.g. phage display), where several attributes of a minimotif are varied in a controlled fashion, thus generating a broad landscape of similar minimotifs with subtle variations [17,18].

To assist the annotator in filling out the form, multiple types of support are provided. Double-clicking on any entry field in the form will display a context menu that gives the suggested choices based on relevant content in the MnM database. In the Modification tab, selecting a modification from the context menu will populate a different field in the form with a PSI-MOD accession number. The Abstract Viewer (Fig. 3A) automatically displays the PubMed abstract of paper that has been selected and highlights keywords and terms in different colors based on attribute entries in the database. The coloring scheme is minimotif (purple), activity (blue), target (orange), putative minimotif (red), affinity (yellow), protein domain (green); if the word "motif" is present, it is bolded. Selection of a paper with a right click also opens the abstract on the PubMed web site and a full text version of the paper, if available, in a web browser. This enables efficient access to full text papers and to other NCBI data using the "Links" hyperlink. Linked data of interest to the annotator includes structure and RefSeq accession numbers.

Another component that assists annotators is the Protein Sequence Validation function (Fig. 3B). Once an accession number has been entered, the protein sequence is automatically retrieved from a local version of public databases such as NCBI and displayed in the Protein Sequence Window. Once loaded, the position of the minimotif in the protein sequence is bolded. This ensures that the minimotif is indeed present in the selected protein.

The paper management tools consist of the Paper Browser, Paper Status Window, and Paper Ranking components, which are addressed later. The Paper Browser shown in Fig. 4A can be used to manage millions of papers. The Paper Browser displays metadata about the PubMed abstracts of all papers entered into a table of the MnM database. The metadata includes PubMed ID, authors, affiliation, journal, publication year, comments, tracking status, paper score, title, URL, abstract, and database source. A paper score (discussed later) is used as a default sort parameter, although the entire table can also be sorted by PubMed ID, paper status, PubMed identifier, publication year, or journal using a pulldown menu. Since the table containing papers has more than 120,000 tuples, only the first 1,000 results of any sort are shown.

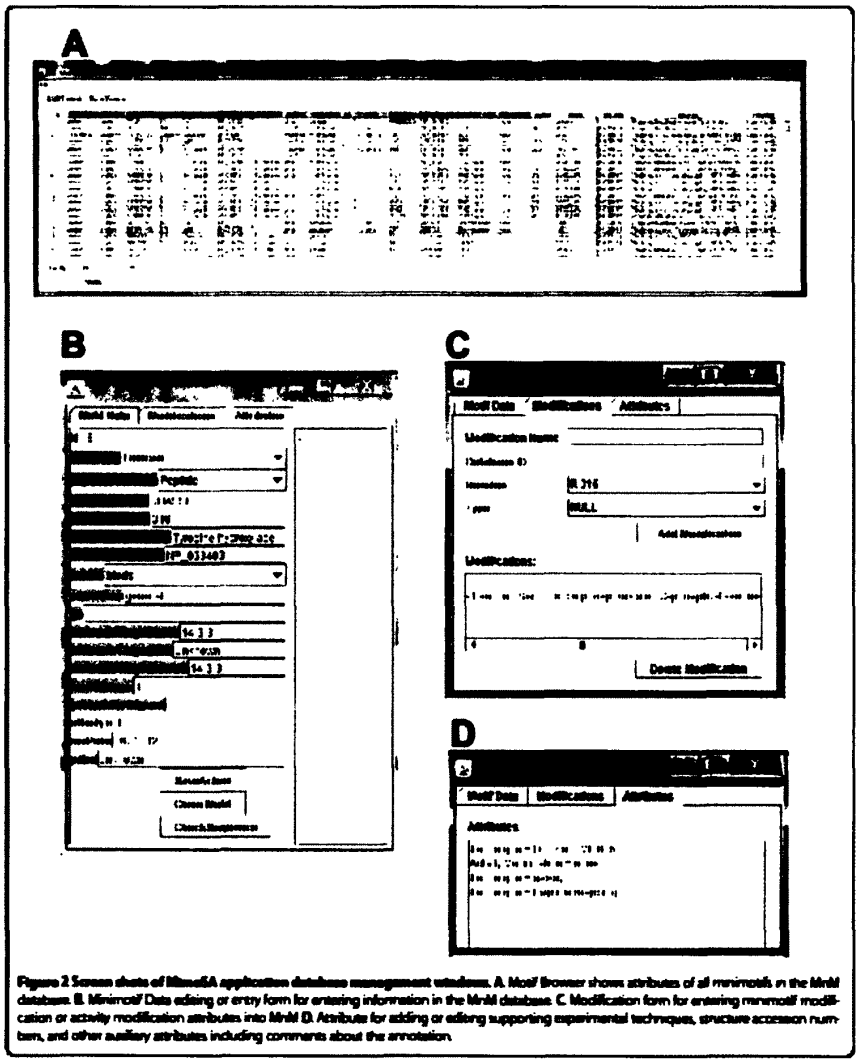


Figure 2 Screen shots of MimoSA application database management windows. A. Motif Browser shows attributes of all minimotifs in the MnM database. B. Minimotif Data editing or entry form for entering information in the MnM database. C. Modification form for entering minimotif modification or activity modification attributes into MnM D. Attributes for adding or editing supporting experimental techniques, structure accession numbers, and other auxiliary attributes including comments about the annotation.

When a PubMed identifier is entered and the "Add Paper" button is selected, the associated paper is retrieved from NCBI and inserted into the database. Any abstract can be retrieved for review by selecting the "Launch by PubMed ID.".

The Paper Status Window, a subcomponent of the Abstract Window, is used to track the annotation status of papers (Fig. 4B). Each time a paper is reviewed and the user updates the status of the paper, a "review event" is created and appended to the paper's history, which is

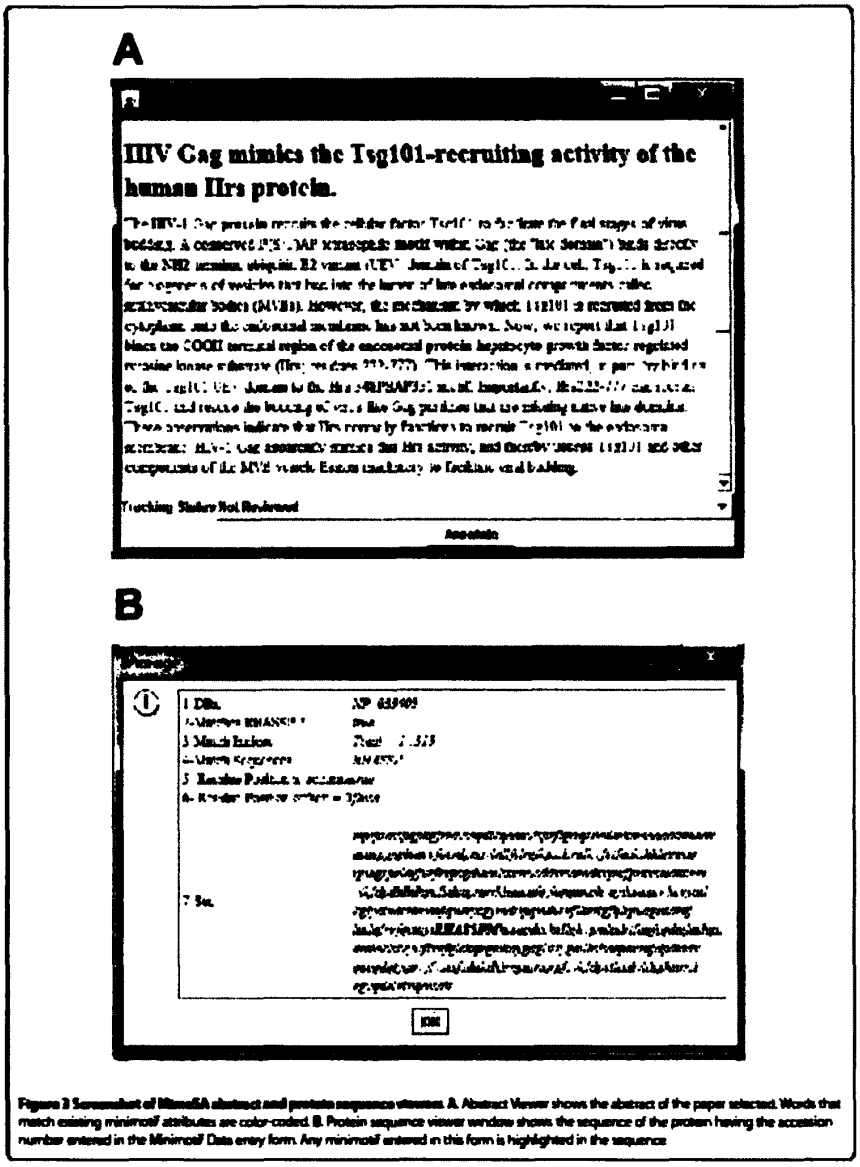


Figure 3 Screenshot of MimoSA abstract and protein sequence viewers. A. Abstract Viewer shows the abstract of the paper selected. Words that match existing minimotif attributes are color-coded. B. Protein sequence viewer window shows the sequence of the protein having the accession number entered in the Minimotif Data entry form. Any minimotif entered in this form is highlighted in the sequence

which contribute to minimotif definitions may either use peptides or full length proteins. We think it is important to specify this as an attribute since the two sources represent very different chemical entities. Finally, we have started using PSI-MOD and GO controlled vocabularies for indicating activities and post-translational modifications of minimotifs.

Identification of papers with minimotif content

The MnM database contains many papers that were previously annotated for minimotif content, but many more papers have yet to be annotated. PubMed contains well over 19 million abstracts of scientific papers. Only those papers that have minimotif content are useful for annotation. Our first approach to pare down the paper list used keyword searches to identify papers, which were likely to contain minimotif content; however, this approach was not efficient. Therefore, we developed new strategies and an efficiency metric for the evaluation and comparison of these strategies (see Additional File 1).

We initially evaluated six general strategies: Keywords/Medical Subject Headings (MeSH), date restriction, forward and reverse citations, authors with affiliations, and minimotif regular expressions. A detailed description of the strategies and results are presented in Additional File 1. These strategies were evaluated using a Minimotif Identification Efficiency (MIE) score, which is defined as the percentage of papers that contain minimotifs. Collectively, these strategies provided a list of approximately 120,000 abstracts, of which ~30% were expected to contain minimotifs based on extrapolation.

Design and training of the TextMine algorithm that scores papers for minimotif content

We wanted to score and rank these papers as a means to better identify the ~30% that contain minimotifs and develop a strategy for scoring all PubMed papers that can be used for future maintenance of the MnM database. To rank papers for minimotif content, we designed the Paper Scoring (PS) algorithm and trained the algorithm using structured data for defined paper sets in the MnM database.

The basic problem of interest can be stated as follows: given a research article (or an abstract), automatically rank the article by its likelihood of containing a minimotif. We used a subset of papers as a training set for training the PS algorithm. Each article in a research article collection A, which is used for training, is read by hand and given a score of either 0, indicating that the paper does not contain minimotifs, or 1, indicating that the paper has at least one minimotif. A similar algorithm has been employed to characterize unknown microorganisms [19]. A crucial difference between the PS algorithm and that of Goh, *et al.*, is that the PS algorithm provides an ordering of the papers instead of using a filter threshold.

The workflow for this phase consists of the following steps: We start with disjoint sets P, N, and T of abstracts, which are positive, negative, or not reviewed for minimotif content, respectively. Let W be the ordered term vector found by taking all significant words (e.g. words like "the", "of", "new" etc., that have no discriminatory value between P and N) from the documents of sets P *and* N. For each word w in W and each article a in P we divide the number of instances of w by the size of a: this is the enrichment of w in a. Then, we sum these enrichments over all P and divide by the size of P to obtain an overall enrichment of w. We repeat this over set N, and subtract the result from w_P to arrive upon a "score" for word w, which ranges from -1 to 1. Higher values indicate more positive association with minimotif content. We now have a vector of decimal "scores", which has the same dimension as W, with one entry per term in the term vector. Call this vector S.

Now, we compute a score for each unknown paper by combining word scores. This phase consists of the following steps.

1) Scan through the paper (or abstract) to count how many times each word w of W occurs in this article.

2) Construct a vector v of all values from (1) in which the order corresponds with S.

3) Compute the correlation between v and S and obtain a Pearson's correlation coefficient pc for each paper. If X and Y are any two random variables, then the Pearson's correlation coefficient between X and Y is computed as $\frac{E(X-\mu_X)E(Y-\mu_Y)}{\sigma_X\sigma_Y}$ where μ_X is the expected value of X, μ_Y is the expected value of Y, σ_X is the standard deviation of X, and σ_Y is the standard deviation of Y.

4) Thus, we have now computed a "score" of the article, which is the Pearson's correlation coefficient between the scored words from the training set W and respective enrichments of those words in the article n.

The Paper Scoring (PS) algorithm's pseudo code is provided in the Additional File 1. The correlation coefficients for the lexemes range from -1.000 to 1.000. This score positively correlates with the presence of minimotif content, as expected.

Paper ranking and evaluation of the paper scoring algorithm

The algorithm above is packaged as an independent application, TextMine, which can be used in conjunction with MismoSA (or as a standalone open source java application which can be integrated with any annotation or analysis pipeline). For the test set, we selected 91 new articles, which we determined to either have or not have minimotif content and were disjoint from the training sets. The basis for all testing of the TextMine application was derived from correlations of TextMine scores to this set.

Vyas et al. BMC Bioinformatics 2010, 11:328
http://www.biomedcentral.com/1471-2105/11/328

The TextMine website and package provides a test data set which reproduces our analysis for a set of test papers. The current version of MimoSA, utilized for MnM annotation, uses scores from TextMine calculated for 120,000 abstracts for paper sorting.

Paper scoring algorithm and training set size

Since the purpose of the algorithm is not simply to rank papers, but rather, to rank papers with increasing sensitivity over time, we evaluated the increase in the algorithms efficacy with respect to larger training sets. We found that there was a degree of variation depending on training set sizes, but that overall, both positive and negative training elements improved the performance (Table 2).

For use in testing TextMine's performance relative to the size of the training set the application package includes an iteration module, which allows for specification of the sizes of positive and negative training sets (this iteration package generated the data in Table 2). We recorded the performance for incrementally increased training set sizes, and noted that as the number of either positive or negative training documents increased, a modest performance improvement was observed. The performance of the algorithm is determined by the correlation coefficient between the calculated scores, between -1 and 1, and an actual score, between 0 and 1.

The table indicates that large increases in the number of positive training articles were comparable to small increases in the number of negative training articles, ultimately showing that both had modest increases in value with set size. A positive correlation coefficient between positive or negative training size and the algorithm performance was observed (0.52 and 0.46, respectively). The correlation score between TextMine scores and the training set scores showed modest increases with size (ranging from 0.59 to 0.66 when using 40 negative and 400 positive abstracts).

The Receiver Operator Characteristic (ROC) curve is a standard metric for visualizing the sensitivity and specificity of an algorithm, which differentiates two populations. We have also included a ROC curve for the highest scoring training set, which had 400 positive and 40 negative articles. We found that this proportion was not required, and that significant correlations could also be obtained with smaller data sizes, as previously described. This curve is shown in Fig. 5. Notably, the area under the curve was above 0.89, indicating a high correlation between the score magnitude and the presence (1) or absence (0) of a minimotif. This data can be generated using the TextMine package. The steps for reproducing this data are described in the TextMine application package.

Table 2: Larger training set sizes (negative, positive) modestly improve algorithm performance

Negative Papers	Positive Papers	Paper Score
10	100	0.60
20	100	0.63
30	100	0.63
40	100	0.64
10	200	0.56
20	200	0.59
30	200	0.58
40	200	0.60
10	300	0.60
20	300	0.63
30	300	0.64
40	300	0.66
10	400	0.61
20	400	0.65
30	400	0.66
40	400	0.66

Because the general utility of this algorithm far exceeds the field of minimotif annotation, we have released TextMine as a stand-alone application that is cross-platform and database-independent.

Discussion

We have built an application that facilitates annotation of minimotifs from the primary literature, which we are currently using to populate a more comprehensive MnM minimotif database. The application scores a set of papers for minimotif content. In principle, the TextMine score can be used to score all PubMed abstracts for minimotif content and can be used in the future for maintaining the

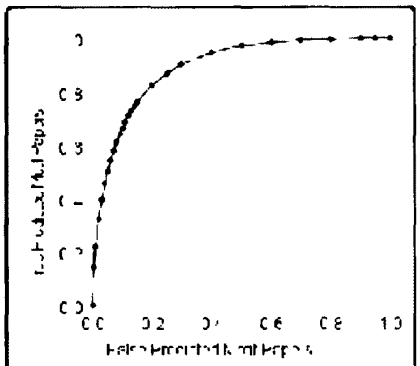


Figure 5 ROC curve analysis of TextMine results. ROC curve as a measurement of the sensitivity and specificity of TextMine for a disjoint test set of 91 pre-scored papers. Area under curve = 0.89

database. As text mining algorithms increase in proficiency and scope, it may be possible to use a large, MimoSA-curated set of minimotif-containing papers as a training set for automatically detecting minimotif definition sentences and phrases in papers by machine learning approaches.

The implementation of the paper scoring algorithm as a SQL stored procedure in MimoSA automates its execution and is amenable to further machine learning development. A static algorithm would have required a word or file list as input and require manual merging of results into the database. One limitation of the TextMine application is that it does not directly control for type biasing. That is, depending on the training set, we expect that there is some risk of "weighting" words heavily to bias previously seen content types. Instead of controlling for this automatically, TextMine outputs the scores of all calculated words so as to enable user inspection of how their training set influences the algorithm. This allows for informed adjustments to the training set on a case-by-case basis.

Although MimoSA was developed primarily for Minimotif annotation, the PS algorithm for scoring content in papers has broader applications. In consideration of its potential use, we have implemented it as a separate program, TextMine. For other annotation purposes, correlation scores for individual words from a training set of articles already known to either contain, or not-contain, the desired information are calculated. This results in a rank order for several thousands of words. For each single article, the PS algorithm then calculates a Pearson's Correlation Coefficient between two large linear sets: the score of each word in the aforementioned dictionary, and the corresponding enrichment of that word in the article's title and abstract. Despite the broad range of semantic methodologies for communication of peptide minimotif information, we still observed significant differentiation of the paper rankings when applied to the minimotif content papers.

Conclusions

The MimoSA application interfaces with a normalized model of minimotif function, facilitating non-redundant annotation of minimotifs. The MimoSA user interface combines a set of features that facilitate annotation, including the browsing, sorting, creation, and modification of minimotif annotation entries. Additionally, interactive paper selection, a database driven Minimotif Annotation Form and literature browser, minimotif attribute based markup and highlighting of abstracts, the display of minimotif positions in protein sequences, and minimotif publication scoring and status tracking. MimoSA also features an adaptive, database-driven paper-ranking strategy that can be used to rank papers for minimotif content, which, when combined with the paper tracking module, represents an adaptive approach to literature scoring and content rating. The layered architecture, generalizable data model of minimotif functionality, and database driven application components enable MimoSA to be readily adapted for other molecular annotation projects.

Availability and Requirements

Project name: Minimotif System for annotation

Project home page: mimosa.bio-toolkit.com, textmine.bio-toolkit.com

Operating system(s): Platform independent

Programming language: Java

Other requirements: MySQL 5.0 or higher, Java Virtual Machine 1.6 or higher,

License: Open Source

Any restrictions to use: This paper must be referenced in any publication that uses MimoSA or TextMine, or any application that is developed based on these core applications.

Additional material

Additional File 1 Additional material. Approach for identifying papers with minimotif content, automated markup of abstracts, and pseudocode for paper scoring algorithm.

Authors' contributions

MRS, JV, and RJN were involved in preparation and editing of the manuscript. TM, SR, VK and MRG were also involved in editing the manuscript. JV, TM, DS, and RJN designed and implemented the software application. SR, MRS, JV and VK were involved in identifying the strategies for paper identification. KK calcu-

listed MIE scores. JV designed and implemented the Paper Scoring algorithm and TextMine application. All authors read and approved the final manuscript.

Acknowledgements
We thank the National Institutes of Health for funding (GM079689, AI078708 to MRS and GM083072 to MRG). We would like to thank members of the Minimotif Miner team for suggestions in preparation of this manuscript.

Author Details
[1]Department of Molecular, Microbial, and Structural Biology, University of Connecticut Health Center, 263 Farmington Ave. Farmington, CT 06030-3305 USA, [2]School of Life Sciences, University of Nevada Las Vegas, 4505 Maryland Pkwy. Las Vegas, NV 89154-4004 USA and [3]Department of Computer Science and Engineering, University of Connecticut, 371 Fairfield Rd, Storrs, CT 06269-2155 USA

Received: 19 February 2010 Accepted: 16 June 2010
Published: 16 June 2010

References

1. Rajasekaran S, Balla S, Gradie P, Gryk MR, Kadaveru K, Kundeti V, Maciejewski MW, Mi T, Rubino N, Vyas J, Schiller MR: **Minimotif miner 2nd release: a database and web system for motif search.** *Nucleic Acids Res* 2009, **37**:D185-D190.
2. Balla S, Thapar V, Luong T, Faghri T, Huang CH, Rajasekaran S, del Campo JJ, Shin JH, Mohler WA, Maciejewski MW, Gryk M, Piccirillo B, Schiller SR, Schiller MR: **Minimotif Miner: a tool for investigating protein function.** *Nat Methods* 2006, **3**:175-177.
3. Gong WM, Zhou DH, Ren YL, Wang YJ, Zuo ZX, Shen YP, Xiao FF, Zhu Q, Hong AL, Zhou X, Gao XL, Li TB: **PepCyber : PPEP: a database of human protein-protein interactions mediated by phosphoprotein-binding domains.** *Nucleic Acids Res* 2008, **36**:D679-D683.
4. Puntervoll P, Linding R, Gemund C, Chabanis-Davidson S, Mattingsdal M, Cameron S, Martin DMA, Ausiello G, Brannetti B, Costantini A, Ferre F, Maselli V, Via A, Cesareni G, Diella F, Superti-Furga G, Wyrwicz L, Ramu C, McGuigan C, Gudavalli R, Letunic I, Bork P, Rychlewski L, Kuster B, Holmer-Ottenich M, Hunter WN, Aasland R, Gibson TJ: **ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins.** *Nucleic Acids Res* 2003, **31**:3625-3630.
5. Obenauer JC, Cantley LC, Yaffe MB: **Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs.** *Nucleic Acids Res* 2003, **31**:3635-3641.
6. Diella F, Gould CM, Chica C, Via A, Gibson TJ: **Phospho.ELM: a database of phosphorylation sites - update 2008.** *Nucleic Acids Res* 2008, **36**:D240-D244.
7. Vyas J, Nowling RJ, Maciejewski MW, Rajasekaran S, Gryk MR, Schiller MR: **A proposed syntax for Minimotif Semantics, version 1.** *Bmc Genomics* 2008, **10**:360.
8. Reeves GA, Talavera D, Thornton JM: **Genome and proteome annotation: organization, interpretation and integration.** *J R Soc Interface* 2009, **6**:129-147.
9. Sherman BT, Huang dW, Tan Q, Guo Y, Bour S, Liu D, Stephens R, Baseler MW, Lane HC, Lempicki RA: **DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis.** *BMC Bioinformatics* 2007, **8**:426.
10. Kawaji H, Hayashizaki Y: **Genome annotation.** *Methods Mol Biol* 2008, **452**:125-139.
11. Reed JL, Famili I, Thiele I, Palsson BO: **Towards multidimensional genome annotation.** *Nature Reviews Genetics* 2006, **7**:130-141.
12. Bracconi QS, Orchard S: **The annotation of both human and mouse kinomes in UniProtKB/Swiss-Prot: one small step in manual annotation, one giant leap for full comprehension of genomes.** *Mol Cell Proteomics* 2008, **7**:1409-1419.
13. Boeckmann B, Blatter MC, Famiglietti L, Hinz U, Lane L, Roechert B, Ulrich A: **Protein variety and functional diversity: Swiss-Prot annotation in its biological context.** *C R Biol* 2005, **328**:882-899.
14. Cohen AM, Hersh WR: **A survey of current work in biomedical text mining.** *Brief Bioinform* 2005, **6**:57-71.
15. Muller HM, Kenny EE, Sternberg PW: **Textpresso: an ontology-based information retrieval and extraction system for biological literature.** *PLoS Biol* 2004, **2**:e309.
16. Gattiker A, Hermida L, Liechti R, Xenarios I, Collin O, Rougemont J, Primig M: **MIMAS 3.0 is a Multiomics Information Management and Annotation System.** *BMC Bioinformatics* 2009, **10**:151.
17. Songyang Z, Shoelson SE, Mcglade J, Olivier P, Pawson T, Bustelo XR, Barbacid M, Sabe H, Hanafusa H, Yi T, Ren R, Baltimore D, Ratnofsky S, Feldman RA, Cantley LC: **Specific Motifs Recognized by the Sh2 Domains of Csk, 3bp2, Fps Fes, Grb-2, Hcp, Shc, Syk, and Vav.** *Mol Cell Biol* 1994, **14**:2777-2785.
18. Kaushansky A, Gordus A, Chang B, Rush J, MacBeath G: **A quantitative study of the recruitment potential of all intracellular tyrosine residues on EGFR, FGFR1 and IGF1R.** *Molecular Biosystems* 2008, **4**:643-653.
19. Goh CS, Gianoulis TA, Liu Y, Li J, Paccanaro A, Lussier YA, Gerstein M: **Integration of curated databases to identify genotype-phenotype associations.** *BMC Genomics* 2006, **7**:257.

doi: 10.1186/1471-2105-11-328
Cite this article as: Vyas *et al.*, MimoSA: a system for minimotif annotation *BMC Bioinformatics* 2010, **11**:328

VENN, a tool for titrating sequence conservation onto protein structure

Published in Nucleic Acids Research, 2009

The VENN application was entirely designed and implemented by Jay Vyas.

Published online 5 August 2009

Nucleic Acids Research, 2009, Vol. 37, No. 18 e124
doi:10.1093/nar/gkp616

VENN, a tool for titrating sequence conservation onto protein structures

Jay Vyas, Michael R. Gryk and Martin R. Schiller*

Department of Molecular, Microbial and Structural Biology, University of Connecticut Health Center, Farmington, CT 06030-3305, USA

Received March 27, 2009; Revised July 1, 2009; Accepted July 8, 2009

ABSTRACT

Residue conservation is an important, established method for inferring protein function, modularity and specificity. It is important to recognize that it is the 3D spatial orientation of residues that drives sequence conservation. Considering this, we have built a new computational tool, VENN that allows researchers to interactively and graphically titrate sequence homology onto surface representations of protein structures. Our proposed titration strategies reveal critical details that are not readily identified using other existing tools. Analyses of a bZIP transcription factor and receptor recognition of Fibroblast Growth Factor using VENN revealed key specificity determinants. Weblink: http://sbtools.uchc.edu/venn/.

INTRODUCTION

In order to gain insight into protein function, scientists often compare orthologous protein sequences (from different species) to identify important residues that are conserved throughout evolution. However, sequences are only a 1D representation of 3D proteins. In this context, it is the spatial configuration of amino acids, not the protein sequence itself, which is under evolutionary pressure. The 3D aspects of the conserved structural motif are not readily decoded from a protein sequence. For example, a binding surface or enzyme active site may have several conserved residues spread over its entire sequence, but in 3D space the residues are consolidated into a localized binding surface.

Many tools such as BLAST have been developed for generating sequence alignments (1). While computational tools such as ConSurf (2) and the Evolutionary Trace Server (3) are very useful to visualize sequence similarity embedded on protein structure, fixed non-interactive selection of similar sequences limits their usefulness. This constraint can obscure details that are critical for understanding proteins and protein families. Here we report VENN, a new program that addresses this limitation. Because it maps the intersect of sequence and structure to evaluate function, it is named after John Venn for his work on Venn diagrams (4).

RESULTS

VENN is a Java application interfaced to a local MySQL database. Users begin by selecting a protein structure, which is retrieved from the Protein Data Bank and displayed using the Jmol molecular viewer (http://www.jmol.org). A BLAST alignment identifies up to 500 putative homologs. Users interactively select among these homologs, and the calculated amino acid conservation at each position is mapped onto the protein structure as a heat map. The application and help videos are at http://sbtools.uchc.edu/venn/.

The VENN workflow is shown in Figure 1A. The user loads the protein structure and sequence into VENN via the Protein Data Bank (PDB) accession number (5). Similar matches to the individual chain sequences (which are putative orthologs or paralogs) in the structure are remotely retrieved from EBI (6) or locally via NCBI using BLAST and stored in the local VENN database. The user selects a set of sequences and initiates an alignment of these filtered sequences, shown in the alignment display. Sequence conservation at each position is calculated from the sequence alignment and used to generate a heat map that is used to color the protein structure in the Jmol structural display window. The user can repeat the filtration process selecting more, fewer, or different groups of sequences to titrate the sequence homology and map it onto the surface of the protein structure. A screen shot of the structural display and alignment windows is shown in Figure 1B.

We have identified four principal strategies for using homolog titration in VENN; users are encouraged to create their own, novel titration protocols: (i) Select all orthologs or paralogs; choosing proteins with the same

*To whom correspondence should be addressed. Tel: +1 702 895 3390; Fax: +1 860 895 3956; Email: martin.schiller@unlv.edu
Present address:
Martin R. Schiller, School of Life Sciences, University of Nevada, Las Vegas, 4505 Maryland Parkway, Las Vegas, NV 89154-4004, USA

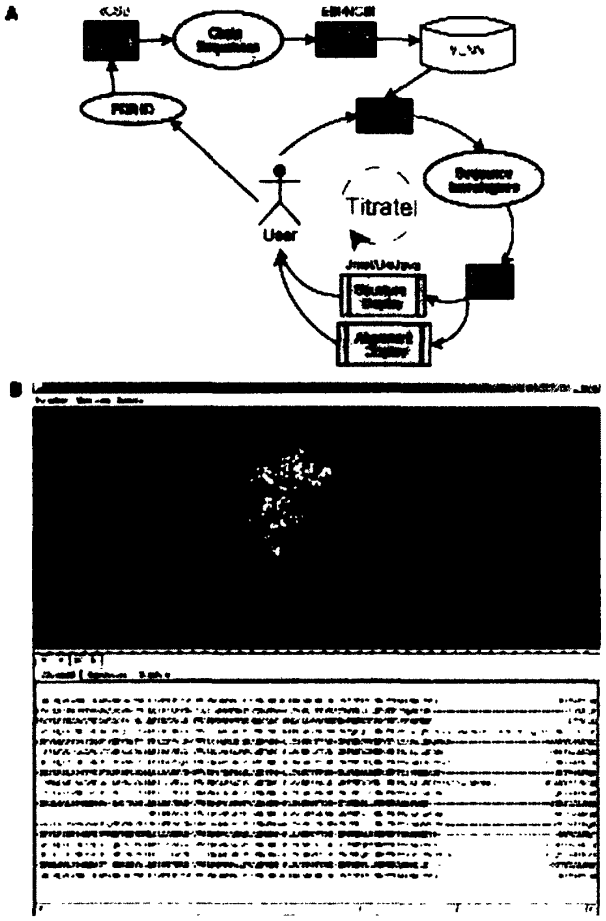


Figure 1. (A) Data processing model for VENN. Processes are shown as boxes (cyan); products are ellipses (orange); displays are yellow. (B) Screen shot of VENN analysed with a complex of Fibroblast Growth Factor 8 (FGF8) bound to a FGF receptor 2c homodimer (blue) (2FDB). Arrow indicates non-conserved specificity residues Thr-Phe in human FGF1,0/4/5/6/7/8/10/11/16/19/20/21/22/23. Residues R1052/G1094/E1131/E1135 in FGF are nearly completely conserved among 15 different FGF family members and contact the FGF receptor.

name can be used for this analysis. This allows a user to determine which regions of the protein are evolutionarily conserved (e.g. Figure 2A); (ii) Select sequences with similar BLAST scores that include different proteins from different species. This reveals important functional sites that are conserved in protein families (e.g. Figure 2B); (iii) Select sparsely distributed sequences with a wide range of BLAST scores. In addition to identifying conserved functional sites in gene families, non-conserved residues can provide clues to the specificity of family members (e.g. Figures 1B and 2C); and (*iv*) Select sequences that have low BLAST scores to reveal the modularity of functional sites in proteins (e.g. Figure 2D).

To demonstrate the utility of VENN we explored these four strategies by examining CCAAT/enhancer-binding protein β (C/EBPβ; PDB: 1GU4), a transcription factor of the bZIP family. The automated BLAST analysis identified 500 C/EBPβ homologs for homology titration. Comparing four orthologs from human, frog, flounder and pufferfish shows high conservation of almost all residues (Figure 2A). As the user titrates in the 50 sequences with the highest BLAST scores representing C/EBP family

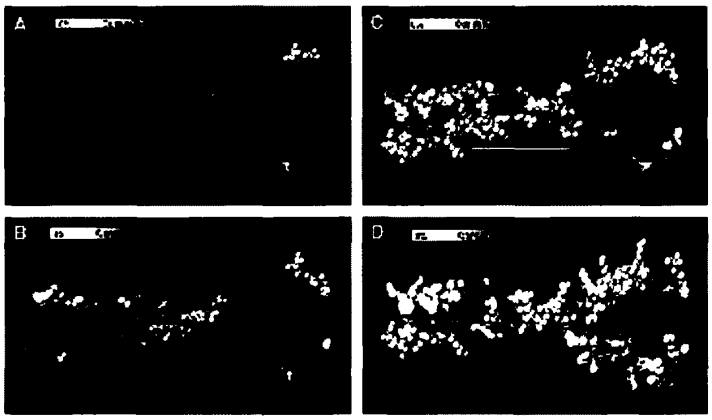


Figure 2. Homology titration of C/EBPβ using VENN. (A–D) Images from VENN analysis of C/EBPβ homodimer (1GU4); chain A is shown using larger spheres. DNA (green) and a heatmap coloring code of residue conservation are shown. Residue conservation maps for putative C/EBPβ homologs are shown: (A) orthologs from four species, (B) homologs with 50 highest BLAST scores, (C) every 20th sequence from the top 160 BLAST scores; inset shows chain A (yellow) with Val 285 (magenta) and chain B (cyan). (D) comparison to coil–coil regions of human myosins and centrosomal protein (290 kDa). Arrows indicate the dimerization interface (cyan) and Val 285 in the DNA-binding site (yellow).

α, β, δ and γ members from many species, functional sites for coil–coil homodimerization and DNA binding emerge (Figure 2B, cyan and yellow arrows, respectively). At the dimerization interface, residues L306, N310, L313, L320, E323, L324 and L327 are completely conserved among distant homologs and form contacts at the dimerization interface. Residues R278, N281, N282, A284, K287, S288, R289 and R295 comprise a DNA-binding site.

To identify differences among closely related members of the bZIP protein family, we selected every 20th sequence in the top 160 BLAST scores (Figure 2C). Within the highly conserved DNA-binding site, V285 (yellow arrow) was poorly conserved. Closer examination reveals that this residue is juxtaposed to a guanine base in the DNA. A literature search revealed that this residue is known to be important for base selectivity in bZIP transcription factors (7). In a similar type of analysis, VENN was used to identify a similar recognition determinant among 15 different FGF family members for binding their receptors (Figure 1B). From this analysis we hypothesize that the critical Thr–Phe residues are specificity determinants for FGF receptor recognition of FGF8 ligands; this was previously recognized for FGF8 isoforms (8).

The BLAST results also revealed several myosins and centrosomal proteins that are not thought to bind DNA, which is supported by a VENN analysis. When the conservation between these proteins is plotted onto the transcription factor, it is clear that the coil–coil dimerization interface remains conserved while the DNA-binding region is not (Figure 2D).

VENN has other unique capabilities. VENN accommodates all protein chains in structures of protein complexes in a single analysis which facilitates analysis of multiprotein complexes. VENN also provides different sequence alignment strategies. A neutral sequence alignment places no weight on any individual amino acid, whereas a BLOSUM alignment weights residues based on the BLOSUM62 matrix (9). VENN also offers a parametric sequence alignment where weights of alignment can be based on the existence of chemical and physical properties of amino acids (for instance, aliphatic, aromatic, acidic, basic, polar). From the visualization perspective, VENN can be used to interactively identify and color regions of protein by searching for a regular expression. Thus, a user could search with 'P.P' to identify any motif that has two prolines separated by one residue. Alternatively, by entering a single amino acid 'M' all methionines can be colored. These features can be used to examine the 3D location of conservation motifs or residues.

DISCUSSION

VENN is an interactive software application that allows users to titrate and map sequence conservation onto protein structures. VENN performs a type of conservation analysis that is distinct from the many programs for pairwise and multiple sequence alignments and from programs such as DALI which is used to identify proteins with similar structural folds (10). Other programs have been published that integrate sequence similarity and protein structure to identify functional sites. VENN is most similar to ConSurf (2), Evolutionary Trace (3) and HomolMapper (11), however VENN has a number of important distinctions that enable new types of discovery. For this section it is helpful to compare an analysis of C/EBPβ with VENN (Figure 2) to that with ConSurf and Evolutionary Trace (Figure 3). HomolMapper has much more limited capabilities.

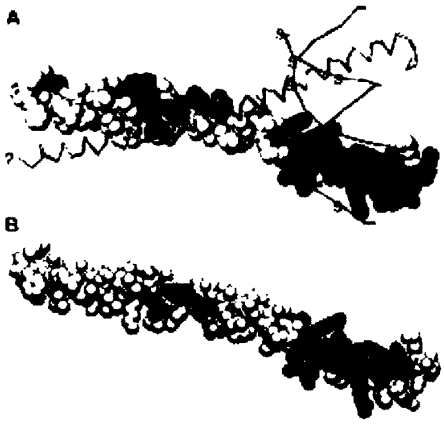


Figure 3. Comparison of ConSurf and Evolutionary Trace analysis of C/EBPβ. (A) Image from ConSurf analysis of C/EBPβ homodimer (1GU4); chain A is shown using larger spheres and chain B backbone is shown; DNA (orange). Color progression from teal to maroon indicates increased conservation; yellow spheres indicate insufficient data. (B) Image from analysis of C/EBPβ homodimer (1GU4) with Evolutionary Trace. Red residues indicate conservation when plotted with the highest Z-score (7.146). Orientations are similar to those for the VENN analysis of the same protein in Figure 2.

Advantages of VENN

VENN has a number of unique features that distinguish it from ConSurf and Evolutionary Trace; however, VENN can be used synergistically with these programs. Most notably, VENN is interactive database-driven program which enables filtration, and iterative selection of different sets of sequences. This is important because it streamlines a number of different strategies for protein sequence selection. Several protein sequence groups can be automatically selected based on species, protein family, motifs, mass, pI, protein length and presence of a user-defined motif. Homologs can also be sorted by BLAST score (default), name, or taxonomy. In order to select different sets of sequences in ConSurf or Evolutionary Trace a user must first perform a multiple sequence alignment and upload a sequence alignment file. This is a limitation: C/EBPβ specificity determinants are only revealed through an ordered interactive titration of homologous sequences (Figure 2C) and in this case not by analyses with ConSurf or Evolutionary Trace (Figure 3). We expect that VENN's flexibility in protein selection and manipulation will enable new types of strategies that we have not yet explored.

VENN also automatically identifies and searches all chains present in a PDB accession number. Therefore, no prior knowledge of the number or identity of chains is required. This user-friendly aspect in VENN is important for exploring multiprotein complexes or complexes of proteins with other molecules. Large structures of complexes, such as nucleosomes, clathrin coats and ribosomes can be analyzed in a single analysis. Often interpretation of a conserved functional site is much easier when visualized in the context of its association with another molecule as exemplified by the conserved residues juxtaposed to a DNA molecule in the structure of the C/EBPβ:DNA complex (Figures 2 and 3). Each chain must be analyzed individually with Evolutionary Trace. While ConSurf can display multiple chains, analysis of multiprotein complexes is slowed by the fact that only one chain at a time can be analyzed.

A number of other features of VENN allow users to readily identify important functional regions in proteins. VENN enables users to select specific residues in the alignment tab; these can be selected and colored on the structure. Motifs can also be selected and colored in the Structure tab; likewise entire domains or protein chains can be colored using either of these functions. Specific residues that are conserved can be identified by examining a multiple sequence alignment in the Alignment tab. Alternatively, holding a mouse over a residue or atom in the structure reveals a popup balloon with its identity. In addition to standard neutral and BLOSUM sequence alignment matrices, VENN also allows flexibility in alignment strategies based on emphasis of different attributes with the aforementioned parametric alignment; e.g. users can heavily weight hydrophobic residues, hydrophilic, etc. ConSurf offers Bayesian or Maximum Likelihood methods for calculating amino acid similarity. By using the Execute Custom Command from the menu a user can enter any Jmol command to modify the display of structure. This flexibility allows users to generate images for publication. While VENN does not have an output function for structure images and alignments, these can be readily captured using a screenshot program (e.g. Snipping tool in VISTA) and the alignments can be cut and pasted into any text editor. VENN can also be used to identify conserved structural features in proteins or proteins families. For example, we used VENN to identify a novel asparagine finger in dynein light chain (1M9L; data not shown) (12).

Synergistic functions in similar software tools

Other tools can be used to complement or precede an analysis with VENN. The ConSurf server, for example, is web based and can be utilized for a quick, automated viewing of highly conserved residues for a single chain in structures of close family members. The Evolutionary trace (ET) program uses a ranking and clustering strategy to map functional sites. Both ConSurf and ET enable more customizable features as well. Other tools, such as SwissPDBViewer (13) and Chimera (14) enable structural modeling and comparison, including alignment of multiple PDB sequences to generate a structural model that relates an entire family of proteins. Such models can serve as novel inputs to VENN for subsequent sequence titration. SwissPDBViewer and Chimera can also be used to manually generate individual figures which resemble those made by VENN by menu and command driven operations. VENN differs from these tools in that it is entirely interactive, integrated with proteome data sources, requires no intermediary file formats for any of its analysis features, and embeds a database and data model of protein sequences/meta data which can directly

automate the aforementioned sequence selection, filtration and titration strategies.

Limitation of mapping homology onto protein structures

VENN, ConSurf and Evolutionary Trace have the major limitation that a protein structure is needed to perform an analysis and there are only ~55000 structures in the latest release of the PDB. One possible solution is to use the ModBase (http://modbase.compbio.ucsf.edu/modbase-cgi/index.cgi) (15) or the Swiss-Model Repository (http://swissmodel.expasy.org/repository/) (16), databases that have millions of structural models that can be downloaded as PDB files. All three programs can read user-defined PDB files. Alternatively, if the query protein is homologous to a protein of known structure, then Swiss-Model can be first used to generate a model structure in PDB file format (17).

CONCLUSIONS

VENN is a novel cross-platform software tool which provides biologists with a highly integrated methodology for visualizing conservation of various functional groups and taxonomical families on the 3D structure of a protein of interest. The ability to readily combine the vast proteomic sequence space with structural information in an automatic fashion can reveal functional attributes which have not been reported using similar tools.

FUNDING

National Institutes of Health (grant numbers GM079689 and AI078708 to M.R.S. and EB001496 to M.R.G.). Funding for open access charge: National Institute of General Medical Sciences, grant number GM079689.

Conflict of interest statement. None declared.

REFERENCES

1 Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25, 3389–3402.

2 Landau,M., Mayrose,I., Rosenberg,Y., Glaser,F., Martz,E., Pupko,T. and Ben Tal,N. (2005) ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res.*, 33, W299–W302.

3 Morgan,D.H., Kristensen,D.M., Mittelman,D. and Lichtarge,O (2006) ET viewer: an application for predicting and visualizing functional sites in protein structures. *Bioinformatics*, 22, 2049–2050.

4 Venn,J. (1880) On the diagrammatic and mechanical representation of propositions and reasonings. *J. Science*, 9, 1–18.

5 Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E (2000) The Protein Data Bank. *Nucleic Acids Res.*, 28, 235–242.

6 Labarga,A., Valentin,F., Anderson,M. and Lopez,R. (2007) Web services at the European bioinformatics institute. *Nucleic Acids Res.*, 35, W6–W11.

7 Fujii,Y., Shimizu,T., Toda,T., Yanagida,M. and Hakoshima,T (2000) Structural basis for the diversity of DNA recognition by bZIP transcription factors. *Nat. Struct. Biol.*, 7, 889–893.

8 Olsen,S.K., Li,J.Y.H., Bromleigh,C., Eliseenkova,A.V., Ibrahimi,O.A., Lao,Z.M., Zhang,F.M., Linhardt,R.J., Joyner,A.L. and Mohammadi,M (2006) Structural basis by which alternative splicing modulates the organizer activity of FGF8 in the brain. *Genes Dev.*, 20, 185–198.

9 Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, 89, 10915–10919.

10 Holm,L.F. and Sander,C (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, 233, 123–138.

11 Rockwell,N.C. and Lagarias,J.C (2007) Flexible mapping of homology onto structure with homolmapper. *Bmc Bioinformatics*, 8, 1–13.

12 Wu,H.W., Maciejewski,M.W., Takebe,S. and King,S.M (2005) Solution structure of the Tctex1 dimer reveals a mechanism for dynein-cargo interactions. *Structure*, 13, 213–223.

13 Guex,N. and Peitsch,M.C (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*, 18, 2714–2723.

14 Pettersen,E.F., Goddard,T.D., Huang,C.C., Couch,G.S., Greenblatt,D.M., Meng,E.C. and Ferrin,T.E (2004) UCSF Chimera - a visualization system for exploratory research and analysis. *J. Comput. Chem.*, 25, 1605–1612.

15 Pieper,U., Eswar,N., Braberg,H., Madhusudhan,M.S., Davis,F.P., Stuart,A.C., Mirkovic,N., Rossi,A., Marti-Renom,M.A., Fiser,A. *et al.* (2004) MODBASE, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res.*, 32, D217–D222.

16 Kopp,J. and Schwede,T (2004) The SWISS-MODEL Repository of annotated three-dimensional protein structure homology models. *Nucleic Acids Res.*, 32, D230–D234.

17 Schwede,T., Kopp,J., Guex,N. and Peitsch,M.C (2003) SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res.*, 31, 3381–3385.

Extremely Variable Conservation of y-Type Small, Acid-Soluble Proteins from Spores of Some Species in the Bacterial Order Bacillales

Published in the Journal of Bacteriology, 2011

The data collection and integration, as well as provisioning of visualization and comparison tools for identifying SSPE's in the entire Firmicute protoeome are presented here. The software necessary for much of this analysis, as well as the data collection, was provided for by Jay Vyas. Experimental work and analysis of phylogenetic data was done in collaboration with the laboratory of Peter Setlow.

JOURNAL OF BACTERIOLOGY, Apr. 2011, p. 1884–1892 Vol. 193, No. 8
0021-9193/11/$12.00 doi:10.1128/JB.00018-11

Extremely Variable Conservation of γ-Type Small, Acid-Soluble Proteins from Spores of Some Species in the Bacterial Order *Bacillales*[▽]

Jay Vyas, Jesse Cox, Barbara Setlow, William H. Coleman, and Peter Setlow*

Department of Molecular, Microbial and Structural Biology, University of Connecticut Health Center, Farmington, Connecticut 06030-3305

Received 5 January 2011/Accepted 28 January 2011

γ-Type small, acid-soluble spore proteins (SASP) are the most abundant proteins in spores of at least some members of the bacterial order *Bacillales*, yet they remain an enigma from both functional and phylogenetic perspectives. Current work has shown that the γ-type SASP or their coding genes (*sspE* genes) are present in most spore-forming members of *Bacillales*, including at least some members of the *Paenibacillus* genus, although they are apparently absent from *Clostridiales* species. We have applied a new method of searching for *sspE* genes, which now appear to also be absent from a clade of *Bacillales* species that includes *Alicyclobacillus acidocaldarius* and *Bacillus suciae*. In addition, no γ-type SASP were found in *A. acidocaldarius* spores, although several of the DNA-binding α/β-type SASP were present. These findings have elucidated the phylogenetic origin of the *sspE* gene, and this may help in determining the precise function of γ-type SASP.

Bacterial spores of species of the *Firmicutes* phylum contain a number of small, acid-soluble proteins (SASP) that comprise 10 to 15% of the protein in the spore's central region or core (30, 32). The following two types of major SASP have been identified in spores: (i) α/β-type SASP that are products of a multi-*ssp* gene family and have extremely similar sequences both within and across species and (ii) γ-type SASP that are almost always encoded by a single *sspE* gene; this is the most abundant protein found in spores of a number of species and comprises 5 to 8% of total spore protein (18–20, 29–32, 39). In contrast to the highly conserved sequences of α/β-type SASP, sequences of γ-type SASP are not well conserved across species, and this has allowed the use of *sspE* and SASP-γ sequences to distinguish closely related *Bacillus* strains and species (17).

In *Bacillus subtilis*, genes encoding both α/β-type and γ-type SASP are transcribed in parallel late in spore development when the various SASP are synthesized, and the transcription of *ssp* genes is mediated by the RNA polymerase sigma factor, σ^G (22). The SASP are degraded soon after spores complete the germination process and begin outgrowth, and one function of these proteins is to serve as a reservoir of amino acids (aa) for protein synthesis early in outgrowth (12, 30, 32). The latter is an important function, since spores become deficient in a number of amino acid biosynthetic enzymes during spore formation and synthesize these enzymes only during spore outgrowth. In addition to serving as a reservoir of amino acids, the α/β-type SASP have an additional function, as these proteins saturate spore DNA and protect it from many types of damage and are thus very important for long-term spore survival (30–32). However, other than serving as an amino acid reservoir, no additional function has been demonstrated for γ-type SASP (12, 30, 32).

In the current work, we have examined genome sequence information for spore-forming *Firmicutes* and have confirmed that (i) spore-forming *Clostridiales* species appear to lack *sspE* genes; (ii) most but not all spore-forming *Bacillales* species, including those in the clade encompassing *Paenibacillus* species, appear to contain a single *sspE* gene; and (iii) some *Bacillales* species, including *Alicyclobacillus acidocaldarius* and *Bacillus suciae*, appear to lack an *sspE* gene. We have also analyzed SASP in spores of several of these species and have found that (i) *Paenibacillus polymyxa* spores contain a γ-type SASP homolog that is related only distantly to γ-type SASP of *Bacillus* species, and (ii) *A. acidocaldarius* spores appear to lack a γ-type SASP but do contain at least two α/β-type SASP. These observations have allowed the determination of the phylogenetic origin of the *sspE* gene in the order *Bacillales*, and this information could lead to suggestions for additional functions of γ-type SASP besides that of an amino acid reservoir.

MATERIALS AND METHODS

Preparation of *P. polymyxa* spores and SASP extraction and analysis. *P. polymyxa* (ATCC 842) was obtained from the American Type Culture Collection. Spores of this species, as well as those of *B. subtilis* PS832, a laboratory derivative of strain 168, were prepared and purified as described previously (14, 23). The purified *P. polymyxa* spores (5 to 6 mg [dry weight]) were lyophilized and dry ruptured with 10-μm glass beads (100 mg [dry weight]) as the alternative, with 10 1-min periods of shaking interspersed with 1-min periods of cooling (23). The dry-ruptured powder was extracted twice with 1 ml cold 5% acetic acid, and the two supernatant fluids were pooled and dialyzed at 4°C in Spectra/Por 3 tubing (molecular weight cutoff, 3,500) for ~18 h against two changes of 1 liter cold 1% acetic acid as described previously (23). Acetic acid extracts of ~4 mg (dry weight) *B. subtilis* spores were prepared as described previously (23) and used as a source of markers for α/β- and γ-type SASP in acid gel electrophoresis. The final dialysates were centrifuged, and the supernatant fractions were lyophilized.

P. polymyxa spores (~12 mg [dry weight]) were incubated in an alkaline decoating solution containing SDS and urea for 30 min at 65°C to remove the spore coat and outer membrane protein while retaining SASP and spore viability (5). After the decoated spores were washed and dried, an aliquot (~5 mg [dry weight]) was disrupted and acetic acid extracts were prepared, processed, and

* Corresponding author. Mailing address: Department of Molecular, Microbial and Structural Biology, University of Connecticut Health Center, Farmington, CT 06030-3305. Phone: (860) 679-2607. Fax: (860) 679-3408. E-mail: setlow@nso2.uchc.edu.

▽ Published ahead of print on 11 February 2011.

dried as described above. Intact *P. polymyxa* spores (6 mg [dry weight]) were also germinated for 60 min at 37°C in ~30 ml of 1 mM dodecylamine in 20 mM Tris-HCl buffer (pH 8.4) (28), and phase-contrast microscopy indicated that ≥90% of the spores had germinated. After centrifugation and lyophilization, ~4 mg (dry weight) of the germinated spores was disrupted and acetic acid extracts were prepared, processed, and dried as described above.

The dried acetic acid extracts from various types of spores were dissolved in 30 μl of 8 M urea plus 15 μl diluent for acid-acrylamide gel electrophoresis, samples were run on acrylamide gel electrophoresis at a low pH, and the gels were stained with Coomassie blue (23). In some experiments, proteins separated by polyacrylamide gel electrophoresis at a low pH were transferred to polyvinylidene difluoride membranes, and the proteins on these membranes were stained with Coomassie blue. Selected stained protein bands were then subjected to automated N-terminal sequence analysis on an Applied Biosystems Procise 494 HT protein sequencer in the Keck Biotechnology Resource Center at the Yale University School of Medicine.

Preparation of *A. acidocaldarius* spores and SASP extraction and analysis. *A. acidocaldarius* strain NRS 1662 was obtained from the American Type Culture Collection, and spores of this strain were prepared on agar plates using a modification of the basal medium for *A. acidocaldarius* as described previously (10). This medium had final concentrations of 1.5 mM $(NH_4)_2SO_4$, 1.7 mM $CaCl_2$, 2 mM $MgSO_4$, 4.4 mM KH_2PO_4, and 1 g/liter yeast extract, and the mixture of these components was adjusted to pH 3.7 with 10 N H_2SO_4 prior to autoclaving and held at 50°C. Autoclaved glucose, agar, and filter-sterilized $MnCl_2$ (all also held at 50°C) were added separately to final concentrations of 1 g/liter, 15 g/liter, and 250 μM, respectively, just prior to being poured into plates. The *A. acidocaldarius* strain was streaked on a plate made as described above except without $MnCl_2$, and the plate was incubated overnight at 55°C. A loop of this overnight culture was inoculated into 20 ml of the medium as described above but without agar, the culture was grown for 5 to 8 h at 55°C to an optical density at 600 nm of ~0.7, and 200-μl aliquots were spread on 40 plates containing $MnCl_2$ as described above. The plates were incubated in bags for ~72 h at 55°C until maximum sporulation had occurred and cooled to 23°C for a few hours, and the cells/spore mix was scraped from the plates and placed in 4°C deionized water. The spores were purified initially by multiple rounds of sonication, followed by centrifugation and water washing as described previously (23), and the crude spores were suspended in 10 ml of 4°C water. Final removal of cells, cell debris, and germinated spores was by layering 1 ml of the crude spores on each of six 13.2-ml ultracentrifuge tubes with 8.5 ml 50% Nycodenz (Sigma Chemical Company, St. Louis, MO) at 23°C and centrifugation at 13,000 rpm in a Beckman 40 Ti rotor for 20 min at 20°C. Under these conditions, cells, cell debris, and germinated spores remained at the water-Nycodenz interface while the pure spores pelleted. The pelleted spores were washed ~5 times with 4°C water to remove Nycodenz and finally suspended in ~5 ml water. This procedure yielded ~45 mg (dry weight) of *A. acidocaldarius* spores that were >98% free of cells, cell debris, and germinated spores as observed by microscopy.

Twelve milligrams (dry weight) of the purified *A. acidocaldarius* spores was dry ruptured and extracted with acetic acid as described above but using 30 min of rupturing and with dialysis for only ~4 h. The dialysate was lyophilized, the residue was dissolved in 25 μl of 8 M urea plus 12.5 μl acid gel diluent (23), 15-μl aliquots were subjected to acid gel electrophoresis, the proteins on the gel were transferred to a polyvinylidene difluoride membrane, the membrane was stained with Coomassie blue as described above, and proteins in stained bands were sequenced as described above.

RESULTS

Sequences of γ-type and α/β-type SASP. The sequences of spores' single γ-type SASP are not as conserved as those of the α/β-type SASP, and γ-type SASP vary from 49 to 139 residues in length (Fig. 1). All γ-type SASP do, however, contain one, two, or three repeats of an evolutionarily conserved 7-aa sequence, TEFASET. In *B. subtilis* and *Bacillus megaterium* SASP-γ, this sequence is the recognition site for cleavage of this protein by a specific protease, termed GPR, that initiates degradation of both α/β-type and γ-type SASP early in spore outgrowth (Fig. 1 and 2, sequences shaded in red) (30, 32). The spacing between the conserved heptapeptide sequences in γ-type SASP that have multiple repeats of this sequence varies considerably (32). In contrast to the α/β-type SASP present in spores of all *Firmicutes* species that have been studied, γ-type SASP and *sspE* genes have not been identified in spore-forming species of the order *Clostridiales* (32).

In order to better understand the genetic history of *sspE*, we aimed to position *Firmicutes* phylogeny with respect to the presence or absence of the *sspE* gene. Computational analysis consisted of two steps: (i) exhaustive, controlled searching of completed *Firmicutes* species genomes for *sspE* orthologs and (ii) construction of a *Firmicutes* species phylogenetic tree based on 16S rRNA sequences. These efforts should phylogenetically position the *sspE* gene to divergence points in *Firmicutes* evolution, thus suggesting critical evolutionary events that led to the emergence of SASP-γ as a protein that facilitates spore outgrowth and might play some additional but unknown role. Exhaustive detection of a homologous gene across multiple genomes is computationally intensive (11). Although tools such as BLAST (2) exist for finding similar protein sequences, the fact that proteins evolve at different rates and by different mechanisms mandates that exhaustive searches for orthologs utilize iteration against different thresholds and alignment methods. We thus constructed an as-yet-unpublished software tool for performing a "multidimensional" sequence search which allows for analysis of all "best hit" proteins across multiple genomes with respect to not just one but a set of several target proteins of interest (the source code can be acquired by contacting the authors directly; the database structure was obtained from reference 35). This tool was applied to *Firmicutes* proteomes to reveal *sspE* genes.

The sequences of γ-type SASP were updated and extracted to a relational database containing over 3 million NCBI proteins from completed proteomes and placed in a smaller data mart containing proteins in the 58 fully curated genomes of *Firmicutes* species (27). The proteome sizes in these 58 species ranged from 2,080 proteins (*Ammonifex degensii*) to 6,238 proteins (*Paenibacillus* sp. Y412MC10 [originally classified as a *Geobacillus* species]). The protein sequences for major sporulation proteins Spo0A, Spo0E, Spo0F, SpoIIE, SpoIIGA, SpoIVB, SpoVAA, SpoVFA, SpoVFB, CwlJ, and CotE from different species were identified and visualized as histograms alongside the best hits when stratified against SASP-γ proteins from *Bacillus clausii*, *B. subtilis*, and *Lysinibacillus sphaericus*, each of which has SASP-γ proteins with rather different amino acid sequences (Fig. 1). Non-trivial best hits (those with a pairwise sequence similarity of 0.3 on a scale from 0 to 1) were investigated for all genomes in order to validate a list of known *sspE*-containing organisms, and for the significant SASP-γ sequence feature, the TEFASET motif was investigated (32), as it appears to be well conserved and thus likely plays a critical functional role for the SASP-γ protein family. The final list obtained represented a complete set (with respect to NCBI-curated *Firmicutes* species genomes) of *sspE*-containing organisms and can be utilized for phylogenetic analysis. Clear *sspE* homologs were found in all *Bacillales* species (Fig. 1) except *A. acidocaldarius*, *B. nuciae*, *P. polymyxa*, *Paenibacillus* sp. IDR-2, *Paenibacillus* sp. Y412MC10 and *Bacillus selenitireducens*. However, *B. selenitireducens* likely does not sporulate and lacks genes for a number of important spore proteins, including α/β-type SASP (6, 24; data not shown). No *sspE* genes were identified in spore-forming *Clostridiales* species either, as was

FIG. 1. Comparison of amino acid sequences of γ-type SASP from spore-forming *Bacillales* species. The sequences are given in the one-letter code for the following species: *Afl*, *Anoxybacillus flavithermus*; Bam, *Bacillus amyloliquefaciens*; Ban, *Bacillus anthracis*; Bat, *Bacillus atrophaeus*; Bav, *Bacillus aminovorans*; Bbr, *Brevibacillus brevis*; Bce, *Bacillus cereus*; Bcl, *Bacillus clausii*; Bcy, *Bacillus cytotoxis*; Bfi, *Bacillus firmus*; Bln, *Bacillus halodurans*; Bli, *Bacillus licheniformis*; Bme, *Bacillus megaterium*; Bmy, *Bacillus mycoides*; Bpf, *Bacillus pseudofirmus*; Bpn, *Bacillus pseudomycoides*; Bpu, *Bacillus pumilus*; Bsu, *Bacillus subtilis*; Bth, *Bacillus thuringiensis*; Bwe, *Bacillus weihenstephanensis* chromosome; Bwp, *B. weihenstephanensis* plasmid; GCS, *Geobacillus* sp. CS6-T3; Gka, *Geobacillus kaustophilus*; Gst, *Geobacillus stearothermophilus*; Gth, *Geobacillus thermodenitrificans*; GWC, *Geobacillus* sp. WCH70; GY1, *Geobacillus* sp. Y412MC61; Pya, *Paenibacillus* sp. Y412MC10; Hha, *Halobacillus halophilus* (originally *Sporosarcina halophila*); Lsp, *Lysinibacillus sphaericus*; Oih, *Oceanobacillus iheyensis*; Pca, *Paenibacillus curdlanolyticus*; Pjd, *Paenibacillus* sp. JDR-2; Pot, *Paenibacillus* sp. oral taxon 786 strain D14; Ppo, *Paenibacillus polymyxa* strain E681; Sur, *Sporosarcina ureae*; and Tth, *Thermoactinomyces thalpophilus*. For species whose abbreviations are underlined, the amino acid sequences were from the *sspE* gene cloned from this species (18, 25, 33); all other sequences were from the species' completed genome sequences. The seven-residue sequences shaded in red are the sites for recognition and cleavage by the SASP-specific protease GPR during spore outgrowth, with cleavage between the bold residues (29, 32). The sequence in the N-terminal region of Ppo in purple is the protein sequence determined in this work, and the yellow blocks of sequences in Ppo and Bsu represent two long blocks of repeated sequences in each of these proteins.

found previously by the analysis of acid-soluble spore proteins or completed genome sequences (7, 8, 32).

Comparison of amino acid sequences of SASP-γ obtained from completed genomes of spore-forming *Firmicutes* species, as well as those identified by targeted gene cloning (Fig. 1) (18, 25, 34), showed that γ-type SASP exhibit some conserved sequences but are most notable for the following characteristics: (i) a significant enrichment of Gln and Asn (8- and 6-fold enriched with respect to normal frequencies in the proteomes of *Paenibacillus* sp. JDR-2 and *Alicyclobacillus acidocaldarius* [~0.036 and 0.105, respectively]); (ii) very low levels of hydrophobic residues, with substantially lower hydrophobic residue abundances in γ-type SASP, although these residues tend to be less naturally abundant in general (e.g., levels of L, I, and P are, respectively, 3, 2.5, and 1.8 standard deviations below proteomic backgrounds); and (iii) 1 to 3 repeats of the relatively well-conserved TEFASET GPR cleavage site, although the spacing between these last conserved sequences varies considerably (Fig. 1). Only a single *sspE* gene was identified in the completed genomes searched, with the exception of *Bacillus weihenstephanensis*, which contained *sspE* genes encoding almost identical proteins on both the chromosome and a plasmid (Fig. 1).

The absence of an obvious *sspE* gene from the completed genomes of *A. acidocaldarius*, *B. nuciae*, *Paenibacillus* sp. Y412MC10, *Paenibacillus* sp. JDR-2, and *P. polymyxa* was surprising, but examination of available draft genomic sequence information for *Paenibacillus curdlanolyticus* and *Paenibacillus* sp. oral taxon 786 strain D14 also revealed no obvious *sspE* genes, although these species as well as *A. acidocaldarius*, *B. nuciae*, *Paenibacillus* sp. Y412MC10, *Paenibacillus*. sp. JDR-2, and *P. polymyxa* did contain multiple genes encoding α/β-type SASP (Fig. 2). The amino acid sequences of α/β-type SASP from *Bacillales* species compared in previous work are extremely well conserved (30, 32). In particular, these proteins exhibit two 18- to 19-aa regions of a highly conserved sequence that are commonly separated by a 3-aa spacer (Fig. 2, red and yellow highlighted regions) and in *B. subtilis* and *Geobacillus*

FIG. 2. Comparison of amino acid sequences of α/β-type SASP from *A. acidocaldarius*, *B. subtilis*, *B. mucilaginosus*, *G. kaustophilus*, and *Paenibacillus* species. The sequences are shown in the one-letter code, with the asterisk in the Bsu3 sequence denoting a stop codon and the dashes introduced to maximize sequence alignments. The sequences are from the following species: Bsu, *B. subtilis*; Gka, *G. kaustophilus*; Aac, *A. acidocaldarius*; Bmu, *B. mucilaginosus*; Pym, *Paenibacillus* sp. Y412MC10; Pcu, *P. cardiovalpinus*; Pjd, *Paenibacillus* sp. JDR-2; Pot, *Paenibacillus* sp. oral taxon 786 strain D14; and Ppo, *P. polymyxa*. The sequences above the break are those from species that are more distantly related to those whose sequences are listed below this break. The long regions of sequence shaded in red and yellow are sequences that each form long α-helices that are important structural elements in these proteins' binding to DNA (16). The site of cleavage of these proteins during spore outgrowth by the SASP-specific protease is between the bold residues in the red-shaded regions. The sequence blocks shaded in green seem to be duplications of the CPFR cleavage site region, and in all of the proteins with this duplication, the spacing between the red and yellow blocks of sequence is greatly increased. The Ppo1, Ppo2, Aac1, and Aac2 sequences in purple were obtained by protein sequence analysis in this work, although the complete sequences of the *P. polymyxa* proteins are from strain D681 and the sequences of the *A. acidocaldarius* proteins are from strain DSM 446.

kaustophilus sequences) (30, 32). These two conserved regions each form long α-helices that interact with the minor groove of DNA and are also a major element of the SASP-SASP dimer interface when the protein is bound to DNA (16). Notably, the spacing between these two conserved regions is 3 aa in almost all of the α/β-type SASP sequences examined to date (32). However, in all but two of the available α/β-type SASP sequences from *Paenibacillus* species, *A. acidocaldarius*, and *B. mucilaginosus*, there are ≥5 aa between these two highly conserved structural elements (Fig. 2 and data not shown), as has also been observed in almost all α/β-type SASP from *Clostridiales* species (32). In addition, even in the two highly conserved regions in the α/β-type SASP from *A. acidocaldarius*, *B. mucilaginosus*, and *Paenibacillus* species there are amino acids present that are not found at these locations in α/β-type SASP from other *Bacillales* species (Fig. 2). These last differences are consistent with the phylogenetically distant relationship between almost all *Bacillales* species and *Paenibacillus* species, *B. mucilaginosus*, and *A. acidocaldarius* (3, 33; see below). It was notable that the amino acid sequences of the four α/β-type SASP from *Paenibacillus* sp. Y412MC10 are rather different from those of most *Bacillales* species, including those of *G. kaustophilus* (Fig. 2) and a number of other *Geobacillus* species (33; data not shown). These data and the presence of an obvious *sspE* gene in all completed genomes of *Geobacillus* species except *Paenibacillus* sp. Y412MC10 are consistent with the recent assignment of *Paenibacillus* sp. Y412MC10 as a *Paenibacillus* species, as it had originally been classified as a *Geobacillus* species (see below).

Analysis of SASP in *P. polymyxa* spores and possible *sspE* genes in *Paenibacillus* species. Although the absence of obvious *sspE* genes from *Paenibacillus* species, *B. mucilaginosus*, and *A. acidocaldarius* was unexpected, the divergence of these organisms from most other *Bacillales* species was an ancient event (3, 33). Consequently, the fact that *Bacillales* species that appeared to lack an *sspE* gene were related only distantly to *sspE*-contain-

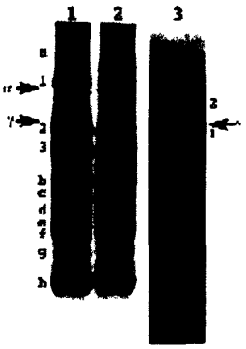


FIG. 3. Polyacrylamide gel electrophoresis at low pH of acetic acid extracts from spores of *P. polymyxa* ATCC 842 (lanes 1 and 2) and *A. acidocaldarius* NRS 1662 (lane 3). *P. polymyxa* spores were isolated and purified, and ~5 mg (dry weight) was disrupted before or after germination; the dry powder was extracted, dialyzed and lyophilized; aliquots were run via polyacrylamide gel electrophoresis at a low pH; and the gel was stained as described in Materials and Methods. The samples run in lanes 1 and 2 are from dormant *P. polymyxa* spores (lane 1) and germinated *P. polymyxa* spores (lane 2). Bands labeled 1 and 3 in lane 1 are the Ppo1 and Ppo2 α/β-type SASP, respectively (Fig. 2), while band 2 is the product of an *sspE*-like gene (Fig. 1), as determined by amino-terminal sequence analysis of these protein bands as described in the text. Bands labeled a to h are ones that were largely or completely removed by decoating treatment, while bands 1 to 3 were not (data not shown). (Lane 3) Dormant *A. acidocaldarius* spores (12 mg [dry weight]) were ruptured and extracted, an aliquot from ~4 mg spores was run via polyacrylamide gel electrophoresis at a low pH, proteins on the gel were transferred to a polyvinylidene difluoride membrane, and the membrane was stained as described in Materials and Methods. Bands labeled 1 and 2 in lane 3 are the *A. acidocaldarius* α/β-type SASP Asc1 and Asc2, respectively, as described in the text. Lanes 1 and 2 are from the same gel, while lane 3 is from a separate gel. The labeled horizontal arrows adjacent to lanes 1 and 3 denote the migration positions of *B. subtilis* SASP-α and -γ, which were determined by running an aliquot of an acetic acid extract of *B. subtilis* spores in lanes that are not shown but were adjacent to lanes 1 and 3.

ing *Bacillales* species suggested the possibility that an *sspE* gene might indeed be present in these distantly related organisms but has diverged sufficiently to preclude recognition by normal sequence comparison programs. Consequently, we examined spores of *A. acidocaldarius* and *P. polymyxa* for a protein that might be an ortholog of SASP-γ.

The acetic acid extract from purified *P. polymyxa* spores produced a large number of bands on polyacrylamide gel electrophoresis at a low pH, with most of these bands migrating faster than *B. subtilis* SASP (Fig. 3, lane 1, *B. subtilis* SASP migration positions denoted by arrows). An obvious question is whether all of the prominent bands seen in the *P. polymyxa* spore extracts were really SASP. To help answer this question, an aliquot of the acetic acid extract from decoated *P. polymyxa* spores was also subjected to polyacrylamide gel electrophoresis at a low pH, and this revealed that decoating did not reduce the intensities of bands 1, 2, and 3 but greatly reduced the intensities of all other bands (data not shown). This suggested that many of the latter bands were due to spore coat proteins. To obtain further evidence that one or more proteins in bands 1, 2, and 3 were indeed SASP, an aliquot of germinated *P. polymyxa* spores was subjected to polyacrylamide gel electrophoresis at a low pH (Fig. 3, lane 2). Bands 1, 2, and 3 were almost completely absent from the germinated spore extract, while the intensities of most other bands were not notably affected. These data are consistent with the proteins in bands 1, 2, and 3 being SASP. Automated N-terminal amino acid sequence analysis of proteins transferred to polyvinylidene difluoride membranes, following polyacrylamide gel electrophoresis at a low pH, produced N-terminal sequences of AQGN NGNS and SRRRNNLQV for bands 1 and 3, respectively. A search of the completed *P. polymyxa* E681 genome indicated that the proteins in these bands were α/β-type SASP, and these were designated Ppo1 and Ppo2, respectively (Fig. 2).

The N-terminal amino acid sequence of the protein in band 2 was PNQGGSXN, and this sequence matched that of a protein, termed Ppo, encoded in the *P. polymyxa* E681 genome (Fig. 1). This protein is clearly not an α/β-type SASP as it lacks the two large blocks of conserved sequence found in these proteins. However, Ppo has a number of similarities with γ-type SASP. In particular, Ppo (i) contains 13% Gln, (ii) has two repeats of a 7-aa sequence with high similarity to the TEFASET motif, where GPR cleaves the *B. subtilis* γ-type SASP (Fig. 1, sequences shaded in red), and (iii) has extended sequence repeats (Fig. 1, yellow blocks in the Ppo and Bsu sequences) (30). In addition, the *ppo* translational start codon was preceded by a strong Gram-positive bacterial ribosome binding site (RBS) and had appropriately spaced −10 and −35 sequences preceding the RBS that were extremely similar to those in promoters of genes encoding highly expressed *Bacillus* sporulation proteins, including SASP that are recognized by RNA polymerase containing σ^G (Fig. 4) (22). Similar RBS and appropriately separated −10 and −35 sequences are also upstream of the coding sequences of the genes encoding the *P. polymyxa* α/β-type SASP Ppo1 and Ppo2 (Fig. 4). In addition, following their translation stop codons, the *ppo*, *ppo1*, and *ppo2* genes each had an inverted repeat sequence, followed by a T-rich region that is likely a rho-independent transcription terminator, as found in all genes encoding major SASP (30).

Surprisingly, the complete genomes of several other *Paenibacillus* species had no strong Ppo homologs (Fig. 1). However, genes encoding proteins with some similarity to γ-type SASP were discovered in *Paenibacillus* species with either completed or draft genomes available by searching for encoded proteins ≤120 aa that matched at least 6 of the 7 amino acids in the TEFASET GPR cleavage motif. To accommodate sequence divergence, we searched via the consensus motif of [AQT]*E.F* [AGS][AST][EQ][FT]. Examination of these potential γ-type SASP revealed at least one ortholog in a number of *Paenibacillus* species for which the coding gene had other features of genes encoding γ-type SASP, including (i) the presence of a strong RBS; (ii) the RBS being preceded by sequences with excellent homology to −10 and −35 promoter sequences recognized by RNA polymerase with σ^G, although these genes could potentially be recognized by the other forespore-specific sigma factor σ^F, as the promoter sequences of σ^G and σ^F-dependent genes overlap to some extent (36); (iii) appropriate spacing between the putative −10 and −35 σ^G recognition elements; and (iv) coding sequences followed by likely rho-independent transcription terminators (Fig. 4). In addition, the

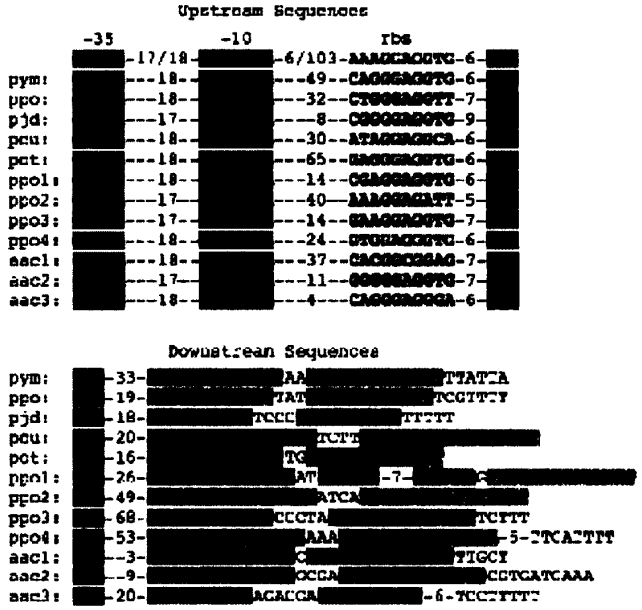
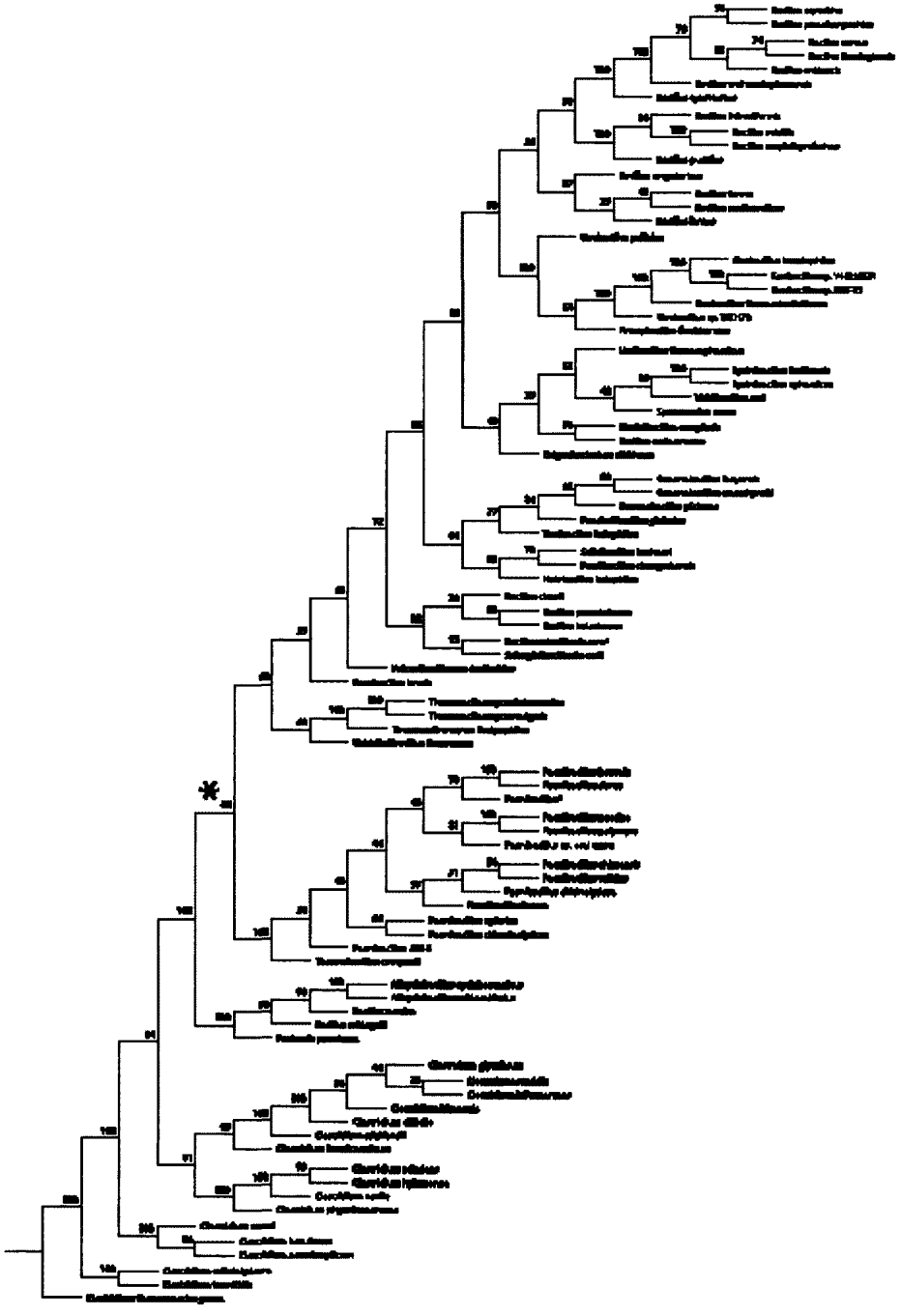


FIG. 4. Putative upstream and downstream regulatory regions for genes encoding α/β-type and γ-type SASP from various species. The genes from the various species are those in Fig. 1 and 2, and the names of the species are as follows: Aac, *A. acidocaldarius*; Bts, *B. subtilis*; Pym, *Paenibacillus* sp. Y412MC10; Pcu, *P. curdlanolyticus*; Pjd, *Paenibacillus* sp. JDR-2; Pct, *Paenibacillus* sp. oral taxon 786 strain D14; and Ppo, *P. polymyxa*. The upstream and downstream sequences of the genes are from the NCBI Microbial Genomes database. The optimal −10 and −35 promoter sequences and their spacing for strong σ^G promoters are listed above the upstream sequences and are from highly expressed σ^G-dependent genes of four *Bacillus* species, including those encoding α/β-type and γ-type SASP (22). Bold nucleotides in the −10 and −35 sequences are >90% conserved in these sequences, and nucleotides that are not bold are 50 to 70% conserved. Note that while the *Bacillus* genes are almost exclusively recognized by σ^G, the recognition sequence for σ^F overlaps that of σ^G to a significant extent (36), so the sequences shown upstream of genes in other species that do or may encode SASP might be recognized by σ^F rather than σ^G. A perfect RBS sequence for *Bacillus* mRNAs is also listed above the upstream sequences. For each gene, in the upstream sequences the putative −10 and −35 σ^G promoter sequences are highlighted in purple, the RBS in yellow, and the translation start codon in red; for the downstream sequences the translation stop codon is highlighted in green and an inverted repeat followed by a T-rich region is in cyan.

putative γ-type SASP from *Paenibacillus* sp. oral taxon 786 strain D14, *Paenibacillus curdlanolyticus*, *Paenibacillus* sp. JDR-2, and *Paenibacillus* sp. YM412MC10 had 9, 12, 14, and 14% Gln, respectively, while the Ppo protein had 13% Gln (Fig. 2). Overall, these data are consistent with the protein encoded by these genes being γ-type SASP, although this has by no means been proven. In addition, the absence of any close homolog of Ppo encoded by other completely sequenced *Paenibacillus* genomes suggests that this group of organisms is extremely diverse, perhaps more so than many other clades of *Bacillales* species.

Analysis of SASP and *sp* genes in *A. acidocaldarius* spores. In contrast to the *P. polymyxa* spore extract, the *A. acidocaldarius* spore extract produced only a single major band (band 1) on acid gel electrophoresis, with this band running slightly faster than *B. subtilis* SASP-γ, as well as at least one other minor band (band 2) (Fig. 3, lane 3, migration position of *B. subtilis* SASP-γ denoted by the arrow). No other band in the *A. acidocaldarius* extract had ≥5% of the intensity of the major band.

Automated sequence analysis of bands 1 and 2 from the *A. acidocaldarius* spore extract resulted in the sequences ANNS GSNR and ANQN[SG]SNR, which were perfect matches to the N-terminal sequence of *A. acidocaldarius* α/β-type SASP Asc1 and Asc2, respectively, with the N-terminal methionine residues removed (Fig. 2). The region upstream of the *asc1* and *asc2* genes contained good matches to likely transcription and translation signals found in other genes encoding major SASP, and both genes had a likely transcription terminator downstream of the translation stop codon (Fig. 4). The identification of only α/β-type SASP in acid extracts of *A. acidocaldarius* spores suggests that if spores of this species contain a γ-type SASP, it is expressed only poorly, is not acid soluble, or is extremely labile to digestion by an acidic protease. However, it seems more likely that this species does not contain an *sspE* gene. Indeed, while a search of the completed *A. acidocaldarius* and closely related *B. suecica* genomes did reveal potential genes encoding proteins that had near matches to the 7-aa GPR cleavage site motif, candidate genes lacked many other features of genes encoding γ-type SASP in either *Bacillus* spe-



1890

cies or *P. polymyxa*, such as a strong RBS and putative σ^{D} promoter sequences (data not shown). Consequently, it appears most likely that *A. acidocaldarius* and *B. succiae* lack an *sspE* gene.

DISCUSSION

The work in this communication indicates that although *sspE* genes are found in most spore-forming *Bacillales* species, they are most likely absent from *A. acidocaldarius* and *B. succiae*, two species that are evolutionarily divergent from other spore-forming *Bacillales* (3, 4, 9, 15, 33, 38; see below). Phylogenetic reconstruction of bacterial evolution can be more straightforward than searches for protein orthologs, and this has been done for *Firmicutes* species in the past (15). We have reconstructed a *Firmicutes* phylogenetic tree that includes all known *sspE*-containing species, along with a number of other *Firmicutes*, including those of *Bacillales* species that lack an *sspE* gene. In order to generate this phylogenetic tree using the most up-to-date information from public databases, we merged 16S rRNA sequences from various prokaryotic genome resources, including the NCBI, RDP, and GreenGenes databases. In particular, the RDP database was used to provide a broad survey of various classes of 16S rRNA sequences to increase the accuracy of our predictions. All reconstructions were performed using the BOSQUE program (26), and critical sequences which appeared to be volatile during alignment were cross-validated via external BLAST searches of EBI prokaryotic 16S rRNA sequences, as well as by comparison with previous reconstructions. Inclusion of sequences of a number of *Clostridium* species was shown to be critical for training the reconstruction to accurately cluster *Thermoactinomyces* species in a manner that was consistent with external BLAST searches of current data, as well as previous reconstructions by other groups. The sequence alignment was done using Muscle 3.6, and the tree was derived using AIC with four categories and PhyML evolution by HKY (1, 13). The final tree (Fig. 5) includes the spore-forming *Bacillales* species with completed genomes, as well as those shown previously to contain an *sspE* gene, even though these species' genomes have not been sequenced (18, 25, 34), as well as members of other *Firmicutes* genera and species. Importantly, members of one such genus, *Thermoactinomyces*, are highly dissimilar to other *Firmicutes* and have very few closely matching neighbors. A BLAST search of 16S rRNA sequences from *Thermoactinomyces* species yielded low scores, indicating that the genus *Thermoactinomyces* may be a phylogenetically coherent group of organisms, as has been suggested in previous analyses (37, 38).

Knowledge of the evolutionary positions of species that do and do not contain *sspE* genes (Fig. 5) now allows us to pinpoint the common ancestor that first acquired the *sspE* genes as between the ancestor of *Paenibacillus* species and that of *Alicyclobacillus* spp./*B. succiae*, since *P. polymyxa* and *Brevibacillus brevis* contain an *sspE* gene, *Thermoactinomyces thalpophilus* contains an *sspE* gene, and its spores contain a γ-type SASP (Fig. 1) (19, 34). This analysis also predicts that *Thermobacillus* spp. will contain an *sspE*-like gene, while *Paneuria* species as well as *Bacillus schlegelii* will lack an *sspE* gene. The determination of at least the *Paneuria penetrans* genome sequence is in progress (21), so a definitive test of this prediction may soon be forthcoming. We note, however, that this overall interpretation assumes that an *sspE* gene did not emerge and was subsequently lost within these taxa. We also do not know how many other potentially informative ancestral taxa are not available for analysis due to *Bacillales* undersampling, and in addition, many *Bacillales* culture collection accessions are misidentified (17).

In addition to the absence of SASP-γ and an *sspE* gene from *A. acidocaldarius* and its closely related species, the absence of *sspE* genes has previously been noted in spore-forming *Clostridiales* species (7, 8, 32). Clearly a γ-type SASP is not essential for spore formation, spore stability, or spore resistance, although SASP-γ does provide an amino acid reserve that can be used in spore outgrowth (12, 30). However, this must not be an essential function, unlike the essential role in spore DNA protection for the α/β-type SASP present in all spore-forming *Firmicutes*. α/β-Type SASP degradation can also supply much amino acid for protein synthesis early in spore outgrowth (12, 29). Presumably, the additional gain in amino acid storage capacity in dormant spores that could be provided by a γ-type SASP does not provide spores with a significant evolutionary advantage, or this is compensated for in other ways. Indeed, at least under laboratory conditions, it is very difficult to demonstrate a major phenotypic effect of loss of SASP-γ from *B. subtilis* spores that contain normal levels of α/β-type SASP (12).

The likely absence of a γ-type SASP from *A. acidocaldarius* spores and the apparent absence of an *sspE* gene from *B. succiae* as well suggest that spores of members of the clade containing these organisms do not contain a γ-type SASP. Perhaps knowing more details about the properties of spores of members of this clade in comparison with spores that do contain γ-type SASP may suggest possible additional functions for this extremely abundant spore protein, other than simply being an amino acid reservoir. In this regard it is perhaps noteworthy that *A. acidocaldarius* is an aerobe. Thus, γ-type SASP seem most likely not to play any significant role in spores' long-term tolerance to oxygen, while this might have been suggested as a possibility had the *sspE* gene appeared only in the transition between the anaerobic *Clostridiales* and the generally aerobic *Bacillales* species.

FIG. 5. Phylogenetic tree for *Firmicutes* species. The tree was constructed using 16S rRNA sequences as described in the text. Organism names in green contain an *sspE* gene, organism names in red do not contain an *sspE* gene, and for organism names in black the completed genome sequence is not available and analysis of the presence or absence of the *sspE* gene has not been carried out. The red asterisk adjacent to *B. selenitireducens* indicates that this species almost certainly does not sporulate, as described in the text. The large green asterisk in the region between the ancestor of the *Paenibacillus* genus and the ancestor of the clade containing *Alicyclobacillus* species indicates the period in *Firmicutes* evolution when the *sspE* gene appeared. The numbers adjacent to interior branch points in the tree are bootstrap values.

ACKNOWLEDGMENTS

This work was supported by a grant from the Army Research Office (P.S.).

We are grateful to S.-H. Park for searching the *P. polymyxa* E681 genome sequence prior to public release for genes encoding proteins whose amino-terminal sequences were determined in this work, to N. Williams and M. Crawford of the Keck Foundation Biotechnology Resource Laboratory at Yale University for performing the protein sequencing, and to M. R. Schiller and T. J. Leighton for helpful suggestions on the manuscript. J. C. Hoch and M. R. Gryk also provided computational resources for the proteome analyses and phylogenetic reconstruction.

REFERENCES

1. Akaike, H. 1974. A new look at the statistical model identification. IEEE Trans. Autom. Contr. 19:716–723.
2. Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403–410.
3. Andersen, J. St., et al. 1999. Phylogenetic analysis of *Paenibacillus pasadenensis* by 16S rRNA gene cloning and sequencing. J. Nematol. 31:319–325.
4. Ash, C., F. G. Priest, and M. D. Collins. 1993. Molecular identification of rRNA group 3 bacilli (Ash, Farrow, Wallbanks and Collins) using a PCR probe test. Proposal for the creation of a new genus *Paenibacillus*. Antonie Van Leeuwenhoek 64:253–260.
5. Bagyan, I., M. Noback, S. Bron, M. Paidhungat, and P. Setlow. 1998. Characterization of *yhcN*, a new forespore-specific gene of *Bacillus subtilis*. Gene 212:179–188.
6. Blum, J. S., A. B. Bindi, J. Buzzelli, J. F. Stolz, and R. S. Oremland. 1998. *Bacillus arsenicoselenatis*, sp. nov., and *Bacillus selenitireducens*, sp. nov.: two haloalkaliphiles from Mono Lake, California that respire oxyanions of selenium and arsenic. Arch. Microbiol. 171:19–30.
7. Cabrera-Martinez, R., J. M. Mesas, S. Setlow, W. M. Watson, and P. Setlow. 1989. Purification and amino acid sequence of two small, acid-soluble proteins from *Clostridium bifermentans* spores. FEMS Microbiol. Lett. 52:139–143.
8. Cabrera-Martinez, R. M., and P. Setlow. 1991. Cloning and nucleotide sequence of three genes coding for small, acid-soluble proteins of *Clostridium perfringens* spores. FEMS Microbiol. Lett. 61:127–131.
9. Charlet, L., et al. 2005. Phylogenetic analysis of *Paenibacillus pasadenensis* by use of multiple genetic loci. J. Bacteriol. 187:5700–5708.
10. Darland, G., and T. D. Brock. 1971. *Bacillus acidocaldarius* sp. nov., an acidophilic thermophilic spore-forming bacterium. J. Gen. Microbiol. 67:9–15.
11. Fabus, D. L., Y. Y. Li, D. J. Arnellius, A. T. Kwon, and W. W. Wasserman. 2006. Improving the specificity of high-throughput ortholog prediction. BMC Bioinformatics 7:270.
12. Hachisuka, H. H., and P. Setlow. 1988. Properties of spores of *Bacillus subtilis* strains which lack the major small, acid-soluble protein. J. Bacteriol. 170: 1403–1404.
13. Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22:160–174.
14. Hsu, Z., et al. 2010. Investigation of factors influencing spore germination of *Paenibacillus polymyxa* ACCC10252 and SQR-21. Appl. Microbiol. Biotechnol. 87:527–536.
15. Hachiya, H., M. Tomita, and A. Kanai. 2007. Evolution of ribonuclease H genes in prokaryotes to avoid inheritance of redundant genes. BMC Evol. Biol. 7:128.
16. Lee, K. S., D. Bumbaca, J. Kosman, P. Setlow, and M. J. Jedrzejas. 2008. Structure of a protein-DNA complex essential for DNA protection in spores of *Bacillus* species. Proc. Natl. Acad. Sci. U. S. A. 105:2806–2811.
17. Leighton, T., and R. Marsh. 2010. Biorepositories and their foundations—microbial forensic considerations, p. 581–601. *In* E. Buckwalter, S. E. Schnatter, R. G. Breese, P. S. Kelm, and S. A. Morse (ed.), Microbial forensics, 2nd ed. Academic Press, Maryland Heights, MO.
18. Lenhart, C. A., et al. 1998. Nucleotide sequence of the sspE genes coding for γ-type small, acid-soluble spore proteins from the round-spore-forming bacteria *Bacillus sphaericus*, *Sporosarcina halophila* and *S. ureae*. Biochim. Biophys. Acta 1396:148–152.
19. Lenhart, C. A., K. R. Finn, H. Setlow, H. F. Fasenow, and P. Setlow. 1986. Cloning and nucleotide sequencing of genes for small, acid-soluble spore proteins of *Bacillus cereus*, *Bacillus stearothermophilus*, and "*Thermoactinomyces thalpophilus*." J. Bacteriol. 167:168–173.
20. Magill, N. G., C. A. Lenhart, and P. Setlow. 1990. Small, acid-soluble, spore proteins and their genes from two species of *Sporosarcina*. FEMS Microbiol. Lett. 69:193–197.
21. Manachini, T. R., et al. 2010. A method for release and multiple strand amplification of small quantities of DNA from endospores of the fastidious bacterium *Paenibacillus peoriae*. Lett. Appl. Microbiol. 50:315–321.
22. Nicholson, W. L., D. Sun, B. Setlow, and P. Setlow. 1989. Promoter specificity of σ^G-containing RNA polymerase from sporulating cells of *Bacillus subtilis*: identification of a group of forespore-specific promoters. J. Bacteriol. 171: 2708–2718.
23. Nicholson, W. L., and P. Setlow. 1990. Sporulation, germination and outgrowth, p. 391–450. *In* C. R. Harwood and S. M. Cutting (ed.), Molecular biological methods for *Bacillus*. John Wiley and Sons, Chichester, United Kingdom.
24. Paredes-Sabja, D., P. Setlow, and M. R. Sarker. 2011. Germination of spores of *Bacillales* and *Clostridiales* species: mechanisms and proteins involved. Trends Microbiol. 19:85–94.
25. Quirk, P. G. 1993. A gene encoding a small, acid-soluble spore protein from alkalophilic *Bacillus firmus* OF4. Gene 128:81–83.
26. Ramírez-Flandes, S., and O. Ulloa. 2008. Bionque: integrated phylogenetic analysis software. Bioinformatics 24:2538–2541.
27. Sayers, E. W., et al. 2010. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 38:D5–D16.
28. Setlow, B., A. E. Cowan, and P. Setlow. 2003. Germination of spores of *Bacillus subtilis* with dodecylamine. J. Appl. Microbiol. 95:637–648.
29. Setlow, P. 1978. Purification and characterization of additional low-molecular weight basic proteins degraded during germination of *Bacillus megaterium* spores. J. Bacteriol. 134:331–340.
30. Setlow, P. 1988. Small acid-soluble, spore proteins of *Bacillus* species: structure, synthesis, genetics, function and degradation. Annu. Rev. Microbiol. 42:319–338.
31. Setlow, P. 2006. Spores of *Bacillus subtilis*: their resistance to radiation, heat and chemicals. J. Appl. Microbiol. 101:514–525.
32. Setlow, P. 2007. I will survive: DNA protection in bacterial spores. Trends Microbiol. 15:172–180.
33. Shida, O., H. Takagi, K. Kawachi, and K. Komagata. 1996. Proposal for two new genera, *Brevibacillus* gen. nov. and *Aneurinibacillus* gen. nov. Int. J. Syst. Bacteriol. 46:939–946.
34. Sun, D., and P. Setlow. 1987. Cloning and nucleotide sequencing of genes for the second type of small, acid-soluble spore proteins of *Bacillus cereus*, *Bacillus stearothermophilus*, and "*Thermoactinomyces thalpophilus*." J. Bacteriol. 169:3088–3093.
35. Vyas, J., et al. 2009. A proposed syntax for Minimotif Semantics, version 1. BMC Genomics 10:360.
36. Wang, S. T., et al. 2006. The forespore line of gene expression in *Bacillus subtilis*. J. Mol. Biol. 358:16–37.
37. Yoon, J.-H., I.-C. Kim, Y.-K. Shin, and Y.-H. Park. 2005. Proposal of the genus *Thermoactinomyces sensu stricto* and three new genera, *Lacisporia*, *Thermopolyspora* and *Seinonella*, on the basis of phenotypic, phylogenetic and chemotaxonomic analyses. Int. J. Syst. Evol. Microbiol. 55:305–400.
38. Yoon, J.-H., and Y. H. Park. 2000. Phylogenetic analysis of the genus *Thermoactinomyces* based on 16S rDNA sequences. Int. J. Syst. Evol. Microbiol. 50:1081–1086.
39. Yoon, K., W. C. Johnson, D. J. Tipper, and P. Setlow. 1981. Comparison of various properties of low-molecular weight proteins from dormant spores of various *Bacillus* species. J. Bacteriol. 146:965–971.

Chapter 5: The R3 Methodology for NMR Structure Calculation in Sparse Data Backgrounds

"The only way of discovering the limits of the possible is to venture a little way past them, into the impossible."

-Clarke's Second Law, From The Journal of Future Studies (Shuck, 2004).

Background

Iterative Structure Calculation

The primary methodologies for structural analysis of proteins include NMR spectroscopy and X-ray crystallography. These techniques are complimentary: X-ray crystallography is effective for determination of "still frames" of large proteins which can be crystallized, whereas NMR is uniquely suited for studying dynamics aspects of protein structures.

Current areas of research include increasing the size limitations of NMR, as well as the level of automation (Gryk et al. 2010, Monetlione et al. 2009). NMR structure calculation using traditional methods requires well-refined, high quality, comprehensive data sets (namely resonance assignments and NOESY peak lists) (Hermann 2002).

Chemical shift assignments and NOESY peaks are the primary input to commonplace methods for structure calculation by NMR (which involve NOESY peak assignment), followed by restraint generation and structure calculation (Hermann et al. 2002). The CYANA noeassign protocol implements an iterative

application of this principal to drastically improve structure quality: back-checking of NOESY derived restraints against previously calculated structures iteratively improves restraint sets. Each such iteration is referred to as a "cycle".

Like any iterative method, noeassign's success depends on a base-case. This case is represented by the initial restraint set and a "seeded" (that is, generated from minimal previous knowledge) structure, which cannot be back-checked against a previous one. This structure generally tends to be a good starting point in "average" or "best" case scenarios (i.e. where data is close to complete), but in non-ideal cases, is less effective (Hermann et al. 2002). In this work, we investigate the substituting of this seeded structure as a mechanism for increasing the accuracy of NMR structure determination from sparse input. Sparseness refers to a relative paucity of empirical data artifacts, namely chemical shifts and peaks, but can also be thought of in terms of restraints, since chemical shifts and peaks are required inputs to the restraint generating algorithms, which are necessary for solving structures, by NMR (Hermann et al. 2002).

Can Bootstrapping Better Guide Iterative Calculations on Sparse Data Sets?

The proposed rationale for a new strategy for structure determination is as follows: *We can increase the number of correct NOESY peak assignments, and thus the accuracy of a structure by improving the initial bootstrap structure.* We present and evaluate a method for accomplishing this, called R3 (Reseed, Recalculate, and Rescue) that is capable of improving or "rescuing" structure

calculations that would otherwise fail (in sparse data backgrounds) by reseeding a structure calculation protocol with a higher quality structure, and then running a noeassign calculation.

The inner-workings of CYANA suggest a natural way to implement R3. We describe this general methodology along with an implementation (for the CYANA program) that is easily adapted. We show that R3 is capable of both success and failure, that is, that variable bootstrapping results discriminate "good" structures from "bad" ones. We also demonstrate efficacy of R3 for increasing the quality of three previously solved proteins of varying sizes (other examples are provided as supplementary data) in sparse conditions. These were simulated over a titration of randomly removed data points in both chemical shift and NOESY space. Finally, we share conclusions regarding several potential applications of this method. There are a number of such future directions, including filtration of structures from a large set of potentially "correct" seeds, increased automation of the NMR calculation workflow, and identification of erroneous peak assignment / restraint artifacts.

Methods

Implementation of the R3 Method

In the R3 methodology we alter the initial structure (which serves as base input to an iterative structure calculation method, such as noeassign). The "noeassign" method can be generically described as follows: Given a set of peaks *p*, chemical shift assignments *c*, angle restraints, and a linear chain of amino acids, we calculate peak assignments *a0* and a corresponding set of

structure restraints *r0*. We then use the set of *r*0 along with initial input data to calculate a protein structure *s*0. After this is completed, we recall p, and recalculate a new set of assignments and restraints (*a*1 and *r*1), as well as a new structure, *s*1. This process is continued a total of n times, until we are satisfied with the structure pN-1. CYANA typically runs 7 of such cycles. To implement R3 in a theoretical sense, we simply substitute *s*0 with a structure obtained by some other means.

Technically, R3 was implemented using the CYANA program via the noeassign macro that operates in an entirely automated fashion. Noeassign scripts were specified in a typical CYANA "CALC.cya" script, which is comparable to the standard noeassign scripts (available at http://www.cyana.org). Specifically, noeassign generates a series of assignments (cycle1.noa), and a structure (cycle1.pdb) from calculated restraints (cycle1.upl). In this case of R3, cycle1.pdb is the "seeded" structure.

Acquisition of Test Data Sets

The chemical shift and peak assignments for proteins were obtained from http://bmrb.wisc.edu, having BMRB ids 15270, 16790, and 6546 (Urlich et al. 2007). The data sets had protein lengths of 111, 128, and 175 residues respectively. Thousands of such data sets are available with differing completeness -- for testing, we used a small subset of well-formatted data sets possesing sufficient data required for structure calculation. To create a broad range of data sets with varying NOESY / chemical shift completeness, 300 experiments were run. For each protein, random chemical shifts and NOESY

peaks were removed for all percentages ranging from 10 to 100. A subtle but important feature of such a pruning is that it removes chemical shifts in each such data set, where the algorithm for pruning is as follows:

0) Define P as the percentage of peaks to retain, and define C as the percentage of chemical shift assignments to retain.

1) Select P% of peaks from each peak list, randomly.

2) Compile all selected peaks into combined peak list, which will be called **p.**

3) Compile all resonance assignments from peaks in **p**, calling this **c** (this is the set of resonances which were assigned to a peak).

4) Remove duplicates from **c.**

5) Randomly select chemical shifts from **c** until C percent of ALL chemical shifts have been selected, or 100% of **c** has been selected. If the number of shifts selected is < than C% of ALL chemical shifts, continue to randomly select new chemical shifts from the remaining chemical shifts (which are outside of the set **c).**

The outputs of (2) and (5) represent pruned peak and chemical shift sets that were used in R3 calculations.

Calculation of Structures (Standard Noeassign and Reseeding)

The results in this work were obtained using Cyana 3 software, but similar results were found when R3 was tested on Cyana 2.1. In the case of R3, 10000 calculation steps were implemented for 7 rounds of iterative assignment and structure calculation. Since each calculation was intensive, they were run in

paralleled on an 8-CPU application server. Particularly long structure calculations for standard noeassign were calculated first as per the noeassign standard methodology. For R3 "rescue" calculations, these calculations were cloned (with exception of all files of cycles 2 through 7, as well as final cyana outputs), cleaned of all cycles (with exception of the first), and reseeded with a cyana structure calculated from a "complete" set of chemical shifts and NOEs (that is, a quality structure calculated from the unfiltered BMRB archive). This can be done by overwriting the "cycle1.pdb" file with a predetermined structure. The Talos+ program was also utilized to generate angle restraints for data sets in all cases. These tasks were automated using the java programming language for data integration in conjunction python scripts for execution.

R3 Evaluation Criteria

To quantify our ability to improve structure calculation by better bootstrapping we report the accuracy (and precision) of all structures. Quantification of accuracy is done in CA RMSDs for simplicity and uniformity of comparison.

Data Analysis

Analysis of the large amounts of data produced by this method was undertaken using scripts that imported cyana data sets using software derived from the VENN application for homology titration (Vyas et al 2009) into a MySQL server.

Results

The performance of the R3 bootstrapping methodology is demonstrated in this section using 600 structure calculations for 3 different proteins retrieved from the BMRB. The BMRB ids for proteins shown in this section are 15270, 16790, and 6546, which have 111, 128, and 175 residues, respectively. To simplify and integrate the discussion of these results we define two calculations: controls and rescues. "Control" structure calculation experiments are calculations that have a standard CYANA seeded structure. "Rescues", in contrast, are those that are seeded with a high quality seeded structure in the first cycle of a CYANA calculation.

We noticed structure improvement on a broad scale for all 300 structures. We define "improvement" as the decrease in CA RMSD between a control calculation and its R3 rescued counterpart. Precision varied more considerably in this approach.

R3's Improves Structure Accuracy over a Broad Range

We first wanted to show that R3 generally increases the quality of structure calculations in sparse data conditions, where chemical shifts and peaks have been removed. Figure 19 shows 6 plots illustrating the general improvements in accuracy obtained in various data sets where 10% to 90% of all chemical shifts or 10 to 90 % of all NOE peaks were removed for each of the three data sets, and Figure 20 visualizes the 3D structure of proteins calculated by the R3 method, as well as the standard CYANA methodology. In each figure only 70% chemical shift and 70% of all peaks were available for calculation of

data. The resulting accuracy and precision is tabulated for all experiments in tabulated in Table 6.

It is clear that R3 becomes begins improving accuracy at 10% data set completeness, all the way up to 90% completeness. The most drastic increase in improvement occurs in the range of 50% to 90% completeness. We found that bootstrapping is not required, nor helpful, when data quality is high enough to resolve the structure independently.

R3 Can Drastically Improve Restraint Set Quality in Sparse Conditions

In extreme scenarios (i.e. when data is sparse), we found that R3 was capable of rescuing structures effectively (up to a certain limit). For example, we note improvements from 8.6 to 2.6 angstroms (BMRB id 15270 calculated with 60% CS and 90% NOE) and 7.6 to 2.94 (BMRB id 16790 calculated with 80% CS, and only 20% NOE). The opposite case was seldom noted (this is visually corroborated by inspecting the 6 plots in Table 6). It is clear that R3 can retain high quality restraints, even in scenarios of extreme sparseness: only 48% and 9% of total restraints were recovered by the R3 method in the last two of the aforementioned data sets, for example. This indicates that a few high-quality restraints can be a very powerful ally in the process of structure determination.

Discussion

R3 could be thought of as improving accuracy in one of two ways. First, it may boost the total number of restraints, so that proteins are more likely to be constrained into a natural conformation. Another explanation for R3's improvements is that it improves overall quality of NOE assignment derived

restraints. These experiments show the latter to be true: When comparing control and rescue calculations, we saw that the overall number of restraints varied only slightly. Our results (Table 6) demonstrate that R3 (as implemented in the CYANA program) is capable of satisfying this requirement: high quality seeded structures cannot rescue a calculation's accuracy in all scenarios. The seeded structure does not "force" convergence to the correct solution in all scenarios, that is, the quality of a structure calculation using the R3 method ultimately is a measurement of the quality of empirically derived input data, and the seed does not appear to bias the results of calculation in a manner which is inconsistent with available peak and chemical shift data.

We might consider several, hybrid approaches to structure determination. Any such method might work by generating an overall protein fold which approximates the correct structure as input to R3. The R3 algorithm could then be applied using a small set of peaks and resonance assignments, with the seeded structure as input. There are many methods which might be put to the task of generating such seeds, including the CS-Rosetta program for structure calculation in the absence of NOESY peaks, the Swiss-Model server for homology based protein structure prediction. Additionally we might be able to refine such seeds by inclusion of restraints from different sources (for example, empirically derived or known restraints on the topology of a protein) might be incorporated into such seeds.

Another intriguing future application for R3 would be as a tool for direct filtration of good structure estimates from poor ones. The synergistic combination of these methods with a technique such as R3 could result in novel, hybrid approaches to structure calculation. Another potential application of this method is in the area of benchmarking and refinement of structure calculations. By randomly removing data and recalculating a structure multiple times we can measure the stability of a structure calculation, with respect to empirical data. Such an analysis could help to determine generic heuristics for structure calculation, while also aiding in the identification of outlier data points in a specific structure calculation attempt.

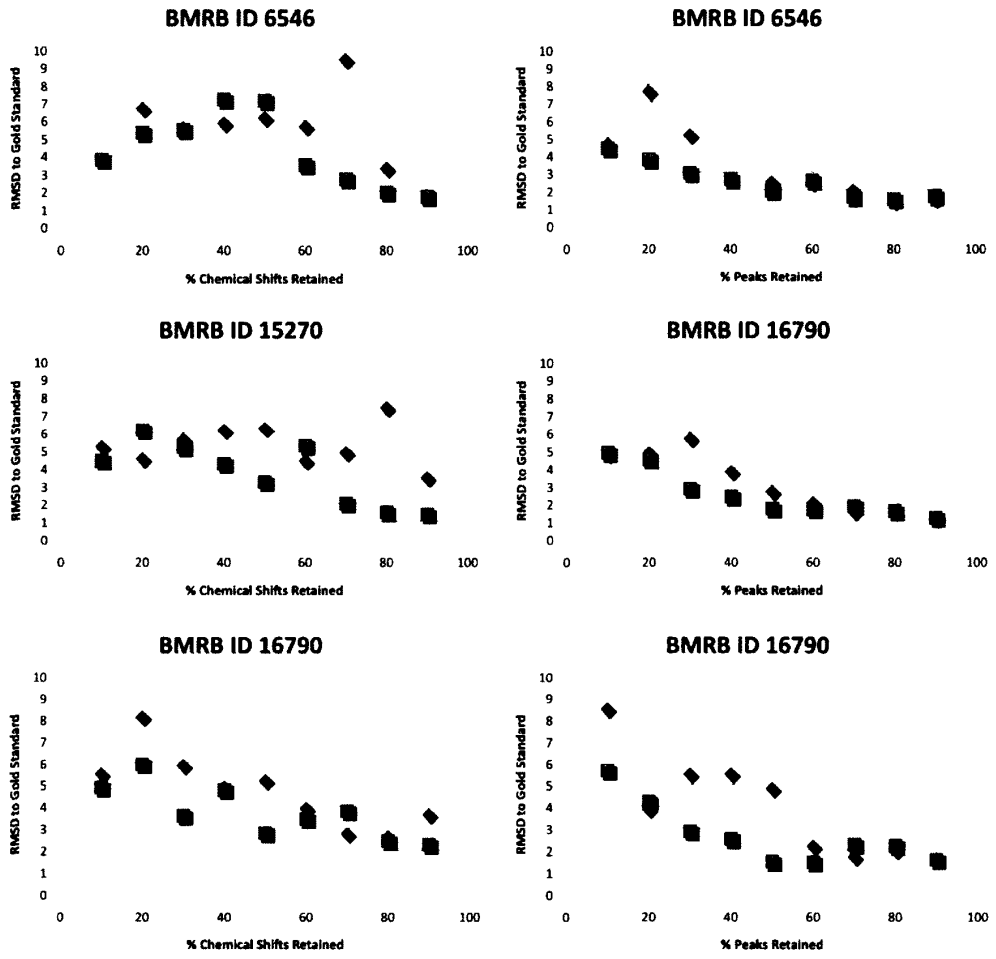


Fig. 18. A comparison of control and R3 calculation accuracy when varying the amount of data for calculation (either chemical shifts or peaks). Red points (diamonds) depict the accuracy (Y-axis) of R3 calculations at a given chemical shift or peak percentage, blue points (squares) represent corresponding control calculation. Accuracy is determined as the CA-RMSD to gold standard structure calculations.

Table 6 (a-f). CA-accuracy and heavy atom precision (reported by CYANA) for 600 structure calculation experiements (continued next page).

a) BMRB ID: 6546
Accuracy

(Percentage of Peaks Retained)

(Percentage of Assignments Retained)

R	10	20	30	40	50	60	70	80	90	**100**
10	7.09	6.83	5.53	5.04	5.26	3.4	3.8	4.82	4.83	4.6
20	3.41	3.91	4.11	5.9	5.18	5.36	7.12	6.21	5.16	7.59
30	3.84	3.72	5.78	5.17	6.11	5.11	5.07	6.26	4.85	5.11
40	4.11	4.21	4.12	5.94	5.05	5.48	4.38	4.83	4.93	2.63
50	5.01	5.37	8.17	5.52	5.22	5.95	5.3	4.36	3.68	2.44
60	6.59	5.07	8.06	8.6	7.21	6.46	5.17	4.35	2.58	2.4
70	3.03	3.68	3.67	5.23	6.21	4.52	4.63	3.88	3.56	1.97
80	6.72	6	6.28	5.73	5.24	11.06	5.47	4.1	4.16	1.35
90	5.02	4.54	5.54	4.48	5.46	6.12	6.12	4.57	2.6	1.5
100	3.76	6.65	5.54	5.81	6.12	5.63	9.4	3.28	1.76	0

(Percentage of Peaks Retained)

(Percentage of Assignments Retained)

C	10	20	30	40	50	60	70	80	90	**100**
10	7.09	6.83	8.12	5.56	4.95	5.7	4.92	2.83	5.62	4.35
20	3.41	4.48	5.03	5.24	4.71	5.37	4.25	5.09	5.37	3.72
30	3.84	5.85	5.53	5.37	5.01	5.9	4.99	3.9	3.31	2.97
40	4.11	3.66	4.83	5.23	7.06	6.41	5.25	2.82	3.45	2.61
50	5.01	4.64	5.88	5.52	5.5	3.73	3.49	3.26	2.33	1.96
60	6.59	5.16	4.29	5.14	6.38	3.36	3.73	2.57	2.52	2.52
70	3.03	5.2	5.37	5.01	5.07	3.71	3.22	2.72	2.47	1.62
80	6.72	5.33	4.86	7.38	6.42	3.64	3.08	1.96	2.09	1.47
90	2.96	5.89	4.73	5.56	4.69	3.03	3.07	1.96	2.51	1.65
100	3.76	5.29	5.46	7.16	7.11	3.47	2.67	1.94	1.68	1.79

b) BMRB ID: 15720
Accuracy

(Percentage of Peaks Retained)

(Percentage of Assignments Retained)

R	10	20	30	40	50	60	70	80	90	**100**
10	6.32	4.9	4.42	7.05	5.54	5.34	4.37	7.83	5.9	4.75
20	5.61	7.35	4.91	6.49	5.31	4.34	5.07	4.93	6.76	4.82
30	3.94	5.65	6.1	4.11	4.28	9.01	5.3	6.73	5.93	5.63
40	6.32	5.01	4.82	5.17	4.43	4.41	4.46	6.1	5.05	3.77
50	4.6	4.44	5.74	6.53	4.89	4.5	5.69	6.11	4.38	2.63
60	4.85	5.8	5.66	6.02	6.52	6.21	6.29	5.1	8.02	2
70	5.78	7.66	6.76	5.44	4.97	5.5	7.8	7.3	4.38	1.52
80	4.94	5.59	4.83	7.02	5.13	5.89	3.99	6.5	2.99	1.5
90	8.99	3.9	5.79	7.07	4.96	8.62	10.21	9.57	3.68	1.13
100	5.18	4.5	5.6	6.12	6.22	4.39	4.85	7.41	3.44	0

(Percentage of Peaks Retained)

(Percentage of Assignments Retained)

C	10	20	30	40	50	60	70	80	90	**100**
10	6.32	4.9	4.42	6.13	3.89	4.95	4.65	4.55	4.4	4.83
20	5.61	7.35	6.82	6.2	5.31	6.28	6.04	5.13	5.69	4.46
30	3.94	5.65	4.27	4.6	4.63	4.85	5.19	5.55	2.84	2.8
40	6.32	2.82	3.62	3.58	5.92	4.54	3.97	3.01	2.14	2.36
50	4.6	6.33	5.26	5.85	4.39	5.62	4.17	3.58	2.05	1.68
60	4.85	5.39	6.61	4.41	5.2	3.35	5.75	2.21	1.63	1.65
70	5.78	3.66	5.09	4.43	8.26	6.16	2.15	2.17	1.49	1.84
80	4.94	5.76	6.16	4.39	3.01	3.23	2.76	1.82	1.38	1.55
90	8.99	4.64	5.1	5.86	3.73	2.59	2.31	2.04	1.71	1.18
100	4.41	6.13	5.17	4.23	3.2	5.25	2	1.5	1.38	1.3

c) BMRB ID: 16790
Accuracy

(Percentage of Peaks Retained)

(Percentage of Assignments Retained)

R	10	20	30	40	50	60	70	80	90	**100**
10	5.26	4.81	6.61	4.52	5.98	5.97	4.57	4.82	5.99	8.43
20	8.25	4.33	3.77	5.54	6.07	4.96	4.39	7.58	4.14	3.85
30	3.79	5.58	5.4	5.63	4.77	9.01	4.57	5.97	4.62	5.46
40	4.93	5.74	7.18	5.52	5.02	4.2	5.97	3.46	3.56	5.47
50	8.25	6.54	4.92	6.29	7.41	5	3.25	4.71	2.94	4.8
60	4.5	5.27	6.17	6.38	4.41	5.33	4.24	5.43	1.97	2.16
70	8.25	4.1	5.96	6.51	6.02	3.88	5.79	4.42	2.92	1.68
80	4.64	5.77	5.11	5.4	4.5	5.21	4.59	1.81	2.5	1.98
90	8.25	4.44	5.2	5.31	4.3	5.1	5.6	4.23	1.66	1.6
100	5.47	8.07	5.85	4.85	5.16	3.89	2.73	2.6	3.61	0

(Percentage of Peaks Retained)

(Percentage of Assignments Retained)

C	10	20	30	40	50	60	70	80	90	**100**
10	5.26	4.81	6.61	3.15	5.3	4.94	5.95	4.91	4.08	5.61
20	8.25	3.79	6.54	4.6	4.7	6.36	6.23	2.94	3.94	4.2
30	3.79	4.67	7.56	4.66	5.3	7.4	4.03	4.97	3.95	2.84
40	4.93	5.6	5.06	5.6	4.17	4.56	4.18	4.56	3.56	2.5
50	8.25	4.74	4.78	5.64	5.3	3.3	4.92	3.25	4.31	1.46
60	4.33	4.29	5.76	5.2	3.13	4.46	5.32	2.65	2.09	1.44
70	8.25	4.63	5.1	6.69	3.71	2.95	3.35	2.53	2.14	2.24
80	4.64	8.83	6.65	5.25	3.43	2.35	5.24	1.83	1.22	2.19
90	8.25	6.66	6.46	4.49	2.88	3.28	5.57	1.91	2.16	1.55
100	4.85	5.91	3.56	4.74	2.76	3.41	3.77	2.41	2.24	1.73

d) BMRB ID: 6546
Precision

(Percentage of Peaks Retained)

(Percentage of Assignments Retained)

R	10	20	30	40	50	60	70	80	90	**100**
10	30.6	28	21.9	22.1	20.9	23.7	18.2	30.6	18.1	14.2
20	28.5	28.8	17.9	21.7	14.6	22.1	13.7	12.1	12.4	7.6
30	30	21.4	29.8	19	13.9	11.8	9.6	6.2	4.8	5.4
40	27.2	32.6	22.1	17.9	11.3	13.6	11.3	4.3	3.5	3.8
50	27.3	25.9	13.1	19.5	11.6	4.7	3.6	2.8	2.7	3.2
60	31	25	21.9	19.8	10.7	8.4	3.8	2.3	2.2	2.4
70	30.3	28.1	21.9	18.4	12.2	4.1	4.2	2.1	1.9	1.5
80	28.4	28.6	20.7	14.1	9.2	5	3.1	2.8	1.8	2.3
90	34.7	26.7	23.2	12.5	5.4	4.4	4.6	1.2	2	1.2
100	28.7	21.6	12.7	10.5	14.6	3.5	2.1	2.8	1.3	1.4

(Percentage of Peaks Retained)

(Percentage of Assignments Retained)

C	10	20	30	40	50	60	70	80	90	**100**
10	30.8	28.2	22.1	22.4	21	23.8	18.6	30.4	18.6	14.7
20	28.5	29	18.2	21.8	15.3	22.4	14.3	12.6	13.1	8.1
30	29.8	21.9	30.1	19.4	14.4	12.4	10.1	6.7	5.4	5.9
40	27.5	32.7	22.3	18.4	11.9	14.2	11.8	4.7	4	4.2
50	27.4	26.2	13.7	19.9	12.2	5.2	4.1	3.3	3.2	3.6
60	30.7	25.1	22.2	20.2	11.2	9.1	4.3	2.8	2.7	2.9
70	30.4	28.3	22.2	18.8	12.8	4.6	4.7	2.5	2.4	2
80	28.6	28.8	21.1	14.6	9.7	5.5	3.6	3.2	2.3	2.8
90	34.9	26.8	23.5	13	6.1	5	5	1.7	2.4	1.7
100	28.6	22	13.3	11	15.2	4	2.6	3.2	1.8	1.9

e) BMRB ID: 15720
Precision

(Percentage of Peaks Retained)

(Percentage of Assignments Retained)

R	10	20	30	40	50	60	70	80	90	100
10	21	23.2	22.4	23.4	22.7	19	17	19	19.8	15.4
20	21.8	22.8	17.5	18.7	19.4	15.7	13	12	13.8	8.8
30	24.3	23.9	21.2	14.3	12.3	14	12.6	9.8	4.3	3.6
40	21	21.9	22.2	14.6	13.7	11.6	9	6.6	3	2.6
50	23.3	16.7	20.1	18.7	12.8	9.5	5.3	2.4	2	1.1
60	21.1	23.3	18.8	14	12.8	8.2	6.3	2.6	0.9	1.2
70	23.1	22.8	20.8	11.2	11.1	8.7	2.1	1.2	0.9	0.6
80	22.4	16.9	15.2	9.1	13.7	3.6	1.8	1.9	1.1	0.8
90	19.9	18.2	18.8	13.5	5.2	3.5	1.8	0.9	0.6	0.5
100	17.1	17.5	12.3	7.6	5.6	3.5	1.6	0.7	0.6	0.5

(Percentage of Peaks Retained)

(Percentage of Assignments Retained)

C	10	20	30	40	50	60	70	80	90	100
10	21.1	22.9	22.4	23.2	22.7	18.4	17.2	19.1	20	15.6
20	21.7	22.7	17.7	18.7	19.4	15.9	13.2	12.3	13.9	9.1
30	24.1	23.9	21.1	14.3	12.5	14.2	13	10.2	4.5	4.1
40	21.1	21.8	22.3	14.8	14.3	11.9	9.6	6.9	3.5	2.8
50	23.3	16.8	20.1	18.7	13	9.7	5.8	2.8	2.5	1.6
60	21.2	23.2	18.8	14.1	13.1	8.3	6.7	2.9	1.4	1.6
70	23.1	22.8	20.9	11.7	11.1	8.9	2.5	1.7	1.3	1.1
80	22.3	17.3	15.4	9.6	14	4.1	2.2	2.5	1.4	1.2
90	20	18.2	18.7	13.8	5.8	4.1	2.1	1.4	1.1	0.9
100	17.3	17.7	12.6	7.9	6	3.9	2.1	1.2	1.1	1

f) BMRB ID: 16790
Precision

(Percentage of Peaks Retained)

(Percentage of Assignments Retained)

R	10	20	30	40	50	60	70	80	90	**100**
10	16.6	14.8	18	16.5	15.2	16.1	13.8	10.5	11.7	13.4
20	16.8	16	13.2	14.6	14.1	10.2	12.3	9.1	8.9	10.6
30	14.9	17.5	15.8	11.3	11.8	11.3	9.6	8.2	3.9	4.5
40	17.1	14.8	12.3	14.4	10.2	6.1	6.6	5.8	4	2.6
50	16.8	17.8	15.8	10.6	10	7.3	6.5	3.7	2	1.4
60	16.9	17.3	11.2	10	4.5	5.4	3.9	2.9	2.7	1.8
70	16.8	11.6	12.7	10.6	5.2	5	2.7	2.1	1.7	2.1
80	14.3	14	10.4	9.5	4.5	3.1	5.2	2.2	1.7	2
90	16.8	14.8	14.6	6	3.9	3.7	3.2	1.6	1.8	1.5
100	16	16.1	11.5	7.4	3.5	4	2	1.6	1.5	1.9

(Percentage of Peaks Retained)

(Percentage of Assignments Retained)

C	10	20	30	40	50	60	70	80	90	**100**
10	16.8	15	18.1	16.6	15.4	16.1	14.1	10.9	11.9	13.9
20	16.8	16.2	13.5	14.9	14.3	10.6	12.5	9.4	9.2	10.8
30	15.1	17.7	15.9	11.7	11.7	11.8	9.7	8.5	4	4.6
40	17.2	14.8	12.6	14.5	10.4	6.4	6.6	5.9	4.3	2.8
50	16.8	17.8	15.9	11	10.6	7.3	6.7	3.9	2.2	1.7
60	17	17.4	11.5	10.2	4.8	5.6	4.1	3	2.8	2
70	16.8	11.8	13.2	11	5.4	5.3	2.9	2.3	1.9	2.2
80	14.5	14.1	10.8	10	4.9	3.4	5.4	2.4	1.9	2.1
90	16.8	15.2	14.7	6.2	4	3.9	3.3	1.8	1.9	1.7
100	16.1	16	11.9	7.7	3.7	4.2	2.2	1.7	1.7	2.1

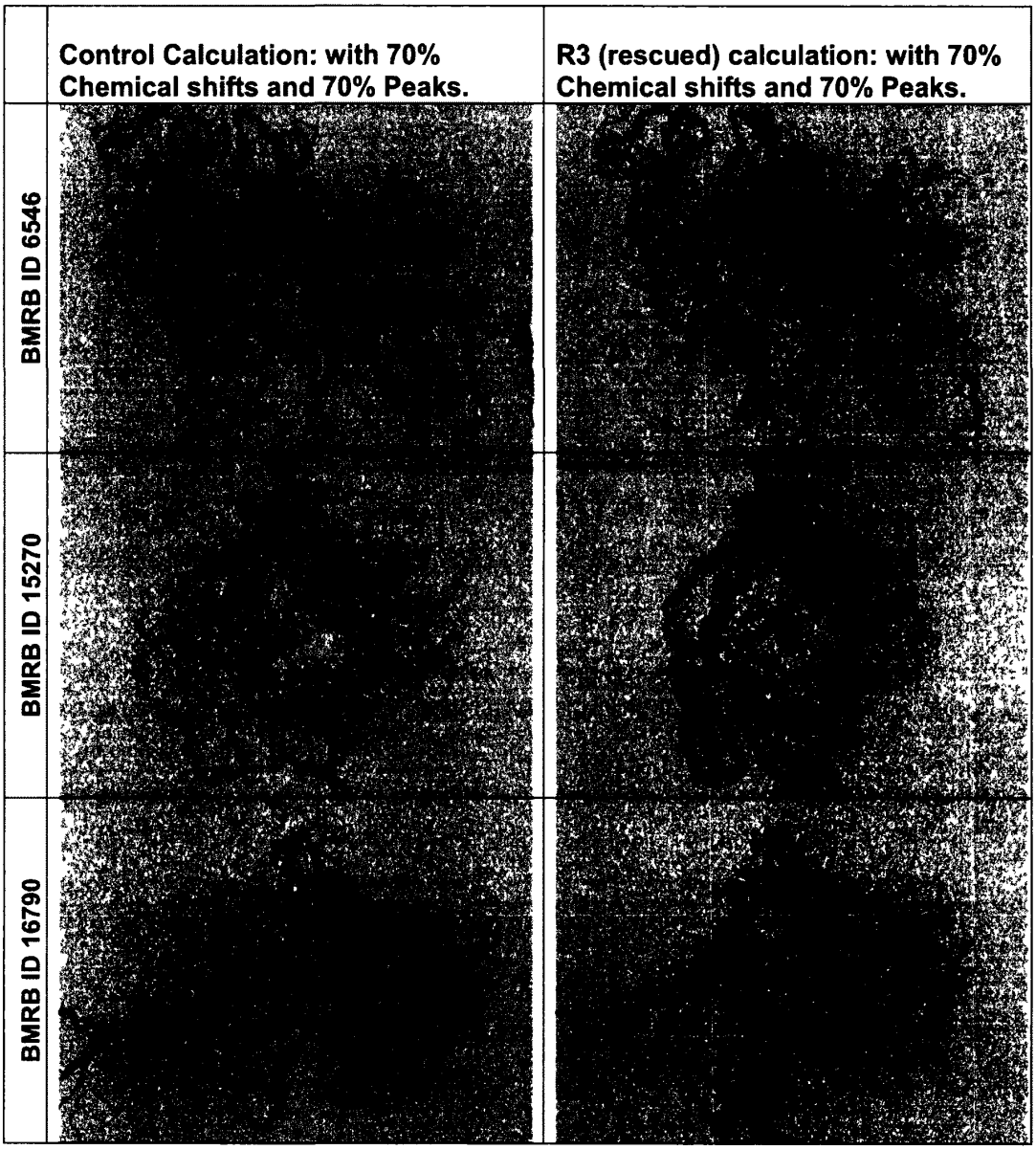


Fig. 19 Control calculations vs rescue calculations visualized. Each row represents a BMRB ID, and the two columns correspond to standard calculations (left) and R3 calculations (right). In red, the correct conformer from the gold standard calculation is shown as reference. In blue, the entire bundle calculated by CYANA for the control (left column) or R3 methodology (right column) is shown. Figures made and aligned with Molmol (Coradi et al. 1996).

Conclusions

"So perhaps the time has come to do some mindless collecting of data."

-Laurie Goodman, From "Hypothesis Limited Research" (1999).

Our interest in integration of protein data originated from a need identified in the "CONNJUR" and "Minimotif-Miner" projects to integrate and normalize data so as to simplify the process of structural and functional protein analysis, respectively. While working towards these ends, it became clear that data integration is a field unto itself. The broad range of content in this thesis supports the notion that data integration is emerging as an increasingly important theme in many areas of the molecular biology of proteins.

Our methods rely on a strategy that is grounded in fundamental information modeling– and we have demonstrated the implementation of this strategy using data-marts, coupled with federated utilities for data ingestion. These technologies facilitate an array of analytical techniques, which can deliver accurate hypotheses to the practicing biologist in an efficient manner.

This work has advanced our understanding of the integrative nature of bioinformatics data in several ways. These pages have provided (1) a robust foundation for defining and mining Minimotif information, (2) a platform for semantically rigorous curation of Minimotifs on a large scale, (3) a practical method for integration of the structural, sequence, and functional aspects of proteins, (4) new insights into the boundaries of the time point in evolution

wherein the SSPE gene emerged in Firmicutes which would be extremely difficult to ascertain without an integrated data processing framework for data mining of bacterial genomes, and (5) exemplary methods for increasing scope of protein structure determination by NMR.

Advances in Minimotif Technology

In the first two chapters of this work I have presented a novel "syntax" defining the information content of Minimotifs that is consistent and unambiguous, and implemented this syntax in a structured relational database. Because this syntax is precise, it was able to be implemented in such a database in a manner that allowed for the querying of various aspects of Minimotif functionality in an intuitive and dynamic fashion.

By coupling a database implementing our model of peptide function to the "Mimosa" application for peptide annotation, we were able to deploy an interactive, high throughput, multi-user technology for the unambiguous annotation of functional peptide motifs. Ongoing work by Schiller and Rajeskeran indicates that this model for molecular function and its ability to be expressed in a structured database can be used to increase the quality of motif searches in the MnM database.

The structured nature of our syntax allows us to leverage the power of large databases and computers for aiding the process of annotation. The ambiguity of common protein annotation vocabularies makes the use of machines as aids in such annotation a less attractive option (machines are notoriously bad at dealing with ambiguous, unstructured data).

Finally, we developed a new algorithm for efficiently discriminating literature abstracts containing data about Minimotifs from other abstracts. This technology is again based on the Mimosa system for annotation, which includes hundreds of thousands of medical abstracts that can be automatically viewed in context of Minimotif related content. This algorithm is generic, and may be applied in other scenarios where the differentiation of text content is desired.

The ability to generically, adaptively rank abstracts could be of much broader useto the research community, and it would be relatively simple to implement an adaptive system which personalized the extraction and clustering of literature for individual investigators. As another, personal offshoot of this technology, I recently deployed the JImpactFactor crawler (http://jimpactfactor.appspot.com) , which has been deployed which outputs all journals which are relevant to a particular area of study, author name, or gene name.

The "sequence" information content of functional of Minimotifs is limited to just a few amino acids in a search string which can potentially match thousands of proteins in a mammalian proteome. However, when we consider the fact that any peptide, in addition to its sequence attributes, contains molecular partners and taxonomical context, it becomes clear that there indeed is more information then sheer sequence at our disposal. In order to utilize these attributes, however, they must be appropriately normalized and modeled. The notion of integrating data to compensate for degeneracies is also a basis for many of the other techniques applied in these pages.

VENN: Bringing "Structural Biology" to Life

The Venn application demonstrates integration that cuts across the many domains of protein informatics – namely sequence, structure, and function, again touching on the theme of broader integration for increasing information resolution. This application, which exhibits broader integration then the Minimotif work, was capable of leveraging our previously constructed API's from the CONNJUR and MnM projects in a synergistic manner.

We are all familiar with the puzzlement posed by a three-dimensional structural model, in spite of its impressive aesthetic qualities. Protein structural interpretation is often difficult because its not always clear which regions of a protein are responsible for which functions – the many biologically inert regions of protein structures (for example, residues at the core) confound our ability to see the signifcance of a protein when viewing it in 3D.

The VENN application allows us to rapidly detect significant substructures in spite of exceeding complexity of 3D coordinates by integrating the latest advances in protein and DNA sequencing to the world of structural biology, allowing the biologist to visualize the consequences of evolution over millions of years in color on a computer desktop. In simple analyses, unimportant residues simply appear white. In more sophisticated workflows, such as that described in the chapter, scientists may identify regions of important function by utilizing sophisticated alignment and "titration" techniques, coupled with careful analysis of residue coloration.

The new version of VENN is web based, allows arbitrary coloring of any sort on the RGB scale, the use of any one of hundreds of alignment matrices available at NCBI, and allows for uploading of custom PDB and fasta-formatted data sets. Venn's exemplary beta-zip transcription factor was just the first step towards development of new paradigm that promises to bring us many returns in the future. We have ultimately taken the VENN system and scaled it into the "HIV-Toolbox" application, which integrates data on an even larger scale (Saergent et al 2011). I certainly envision the continued integration of VENN with more biological data as time goes on – including sequence isolates, DNA sequence conservation for nucleotide bound structures, and gene-ontology terms.

Its quite interesting to consider the consequences of higher throughput structure determination on such applications, since these advances will ultimately increase the data available to tools like VENN by orders of magnitude. At some point in the future, it might be possible that VENN allows for titration of structural as well as sequence changes in a single visual environment.

Data Integration for Distantly Related Proteins

Sequence similarity of genes and proteins is essential for use of common gene finding tools such as BLAST. However, there are cases where a gene's function is not reliant on its primary sequence. In such a scenario, common protein sequence based searches may not readily find true homologs. The SSPE protein in firmicutes is a textbook case of such a gene, which shares very little amino acid sequence similarity to its neighbors.

A desire to find "all" SSPEs in the firmicute proteome inspired the work of Chapter 4. In this work, we demonstrated and defined an entirely new method for sequence scanning and prediction of gene emergence. In particular, to find the sequences, we expanded data regarding genes and their sequence homology into a two dimensional plot of histograms, where, for each particular species, we plotted a row with a histogram visualizing percent similarity of well conserved, poorly conserved, and SSPE proteins. This "controlled" visualization of homology was only possible in context of end-to-end data integration of taxonomical and sequence data into a high-performance data mart.

We have thus integrated the process of sequence mining methods with phylogenetic reconstructions, so as to enable new methods in bacterial protein sequence mining; identifying the phylogenetic origins of the elusive SSPE gene, and shedding light on a particularly interesting time point in bacterial evolution. Current work by Hao and Setlow has since revealed that the origin of the SSPE gene, which was identified in Chapter 4, may be, in fact, a major divergence point the divergence of Firmicute genomes, and thus, in the evolution of microorganisms (personal communication).

An important aspect of this work was our expansion of standard sequences searches into two-dimensional searches, which plot various genes in one. We can envision a powerful alternative to standard BLAST searches based on this paradigm that is not specific to Firmicture proteomes, but rather, which is integrated with the entireity of NCBI's proteomic resources. Such a tool could be useful for gene hunting on a much wider range of species.

CONNJUR: Pushing the Limits of NMR Data Integration

The notion of integration for its own sake is a founding precept of the CONNJUR project that has found its way into every chapter of this thesis, and ultimately, has now become a primary principle of the ongoing works of the MnM project. In Chapter 5, we come full circle to continued advancing technologies that facilitate the NMR workflow for protein structure calculation, which is one of the main goals of the CONNJUR project, by using the integrated strategies which define the overall CONNJUR project in general.

The R3 methodology for structure calculation, although in its infancy, may have implications for higher-throughput structure determination methodologies as well as benchmarking. In addition to ongoing improvements in NMR data processing and analysis, we are advancing our understanding of how amino acid sequences "fold" into three-dimensional structures. We now know that there are a limited number of "folds", based on research done into categorization and clustering of different protein families(Andreeva 2004). Molecular dynamics methodologies will surely benefit from our continually improving understanding of the structural properties of proteins. This in turn will lead to increasingly accurate methods for structural simulation and calculation that rely on such molecular dynamics methodologies for in silico simulation of the protein folding process.

Nevertheless, we will need to validate protein models using empirical data in the future. The fact that we have demonstrated this ability in R3 is thus a proof-of-principal that, as in-silico structural models become increasingly accurate, we may be able to begin solving structures in extremely high

throughput by simply validating these models by collecting a small amount of data.

In a broader sense, R3 is emblematic of the CONNJUR goal – which is the integration of structures, peaks, chemical shifts, and atoms into a pipeline which can be adaptively adjusted, tweaked, iterated, and visualized on the fly with little or no need for manual intervention and file formatting. The prototypical structure calculation models used to automate the R3 experiments represent the first iteration of such a framework for CONNJUR, and ongoing work in the Gryk laboratory continues to "push" the scope of the CONNJUR project to the point where all NMR data types can comingle in a synergistic manner. As a group, the CONNJUR team has also recently released a comprehensive, open source, and vendor neutral spectral data conversion utility to the NMR community, which is the first tool of this sort in the field (Nowling et al. 2011). In an even broader sense, R3 represents a primary goal of this thesis: the demonstration of the fact that integration alone can enable solutions to problems that are otherwise difficult to solve.

The Future of Bioinformatics

This work was not intended to impose a top-down strategy for integrated analysis of protein data on all bioinformaticians, but rather, to explore a broad range of methodologies for integrated analysis of protein sequence, structure, and function in several specific areas, which will generally guide others in the future. To this end we have succeeded. Bioinformatics continues to grow and expand in parallel with improvements made in other related industries – such as

physics, chemistry, and of course, computing. The next several years of bioinformatics promise to be as interesting as any thus far. One particularly interesting trend is the rise of highly efficient methods for analyzing large data sets.

The current climate for data mining is burgeoning with innovation in the area of large-scale data analytics. Recent advances in generic data mining techniques have now affected the trajectory of bioinformatics efforts as well (Taylor et al. 2010). The fast approaching eras of personalized medicine and high-throughput structural biology are destined to increase our data processing requirements by orders of magnitude – while also augmenting our understanding in an equally dramatic fashion.

As one would expect, the pace of progress in information integration is breakneck, and things are changing rapidly. The debate which rages on is not "Should we integrate?" but rather "How should we integrate?" The bioinformatics world stands at a cross roads, where structured data integration techniques, such as those enlisted in these pages, are being challenged by an ambitious and extremely high performance array of "NoSQL" technologies (named after their often cavalier eschewing of traditional SQL-oriented, highly structured database integration technologies).

In the area of data science, these methods value simplicity over explicitness, throughput over precision, and scalability over transactional security. NoSQL technologies have burst onto the bioinformatics scene in the

past 5 years, and are now being applied to the service of protein sequence alignment, genome assembly, literature mining, and even structural biology.

The philosophical basis for these techniques does not solve perennial problems of semantics and data integration, but rather, combats these issues using an entirely lateral method of attack: rather than forcing our data to be correct, allow it to it be incorrect – and simply collect more of it.

Might it be possible, rather than integrating existing, fractionated repositories, to simply recollect biological data on a massive scale and reprocess it using modern, ultra-high performance data analysis technologies? In the biology community, we have seen similar trends in thinking in the area of gene expression analysis and large-scale proteomics. These endeavors, which may be criticized as "noisy" by some, have revolutionized our ability to profile the salient characteristics of a cellular population. Certainly, the CS-Rosetta paradigm, which involves the generation of tens of thousands of candidate protein structures, represents a "big-data" approach to structure determination that, although in its infancy, represents a foreshadowing of things to come. In particular, these paradigms are generally highly dependant on the use parallel computing.

Final Thoughts

Often in science, breakthroughs come in strange, unpredictable forms. The next great advances in biology may very likely come not from larger, more restrictive models of molecular classifications and hierarchies, but rather, from

novel, highly simplified models for dealing with biomolecular computation which have never before been imagined.

Doolittle and others came to witness the importance and power of *sequence-oriented bioinformatics* for evolutionary inference in the last quartile of the 20th century. The next several decades will witness the power of *global bioinformatics data integration* in a similar light. That is, as we improve our ability to integrate computational analysis of proteins, we will witness a deeper conceptual integration of sequence, structure, and function. Ultimately, these will beget a deeper understanding of the combinatorial, expansive molecular relationships that drive cellular function. The curtains are about to rise on the next act of "the greatest show on earth".

Finally, a personal note: As our understanding of molecular interactions continues to improve, we must never forget our prenomial charge – which is the sharing of these advancements with humanity at large. We can do this at the micro-scale by making our software free and open source for all to utilize. Additionally, we may do this on a global scale by continuing to promote bioinformatics to the status of a first-class, exhibitionary science. I cannot imagine that the delicately crafted nuances of protein sequence alignment, the pleasures of virtually spinning large, DNA bound protein models, and the extreme diversity of natures protein arsenal is of interest only to the bioinformatics community. After all, the mysteries of bioinformatics are but a reflection of the regular ongoings that are native to all living things.

References

Aasland R, Abrams C, Ampe C, Ball LJ, Bedford MT, Cesareni G, Gimona M, Hurley JH, Jarchau T, Lehto VP, Lemmon MA, Linding R, Mayer BJ, Nagai M, Sudol M, Walter U, Winder SJ: **Normalization of nomenclature for peptide motifs as ligands of modular protein domains**. *FEBS Lett* 2002, **513:**141-144.

Andreeva, A. 2004. **SCOP database in 2004: refinements integrate structure and sequence family data.** *Nucleic Acids Research* 32, no. 90001: 226D-229. doi:10.1093/nar/gkh039.

Anfinsen CB: **Principles that govern the folding of protein chains.** *Science.* 1973, **96**:223-30

Akaike H: **A new look at the statistical model identification.** *IEEE Trans Automat Contr* 1974, **19**:716-723.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**:3389-3402.

Anderson JM, Preston JF, Dickson DW, Hewlett TE, Williams NH, Maruniak JE: **Phylogenetic analysis of *Pasteuria penetrans* by 16S rRNA gene cloning and sequencing.** *J Nematol* 1999, **31**:319-325.

Ash C, Priest FG, Collins MD: **Molecular identification of rRNA group 3 bacilli (Ash, Farrow, Wallbanks and Collins) using a PCR probe test. Proposal for the creation of a new genus *Paenibacillus*.** *Ant Van Leeuwen* 1993-1994, **64**:253-260.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology**. *Nat Genet* 2000, **25:**25-29.

Aubry M, Monnier A, Chicault C, de Tayrac M, Galibert M, Burgun A, Mosser J: **Combining evidence, biomedical literature and statistical dependence: new insights for functional annotation of gene sets**. *BMC Bioinformatics* 2006, 7:241.

Bagyan I, Noback M, Bron S, Paidhungat M, Setlow P: **Characterization of *yhcN*, a new forespore-specific gene of *Bacillus subtilis*.** *Gene* 1998, **212**:179-188.

Balla S, Thapar V, Luong T, Faghri T, Huang CH, Rajasekaran S, del Campo JJ, Shin JH, Mohler WA, Maciejewski MW, Gryk M, Piccirillo B, Schiller SR, Schiller MR: **Minimotif Miner, a tool for investigating protein function**. *Nat Methods* 2006, **3:**175-177.

Barker WC, Dayhoff MO: **Viral src gene products are related to the catalytic chain of mammalian cAMP-dependent protein kinase**. *P Natl Acad Sci USA* 1982, **79**:2836-2839.

Barton GJ: **Scop: structural classification of proteins**. *Trends Biochem Sci* 1994, **19:**554-555.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The protein data bank**. *Nucleic Acids Res* 2000, **28:**235-242.

Bernstein FC, Koetzle T, Williams G, Meyer E, Brice M, Rodgers J, Kennard O, Shimanouchi T, Tasumi,M: **The protein data bank: a computer-based archival file for macro-molecular structure.** *J Mol Biol 1977*, 112: 535–542.

Bergeron B: **Bioinformatics Computing.** *Pearson Education, Inc.* 2003, ISBN:0-13-100825.

Blum JS, Bindi AB, Buzzelli J, Stolz JF, Oremland RS: ***Bacillus arsenicoselenatis*, sp. nov., and *Bacillus selenitireducens*, sp. nov.: two haloalkaliphiles from Mono Lake, California that respire oxyanions of selenium and arsenic.** *Arch Microbiol* 1998, **171**:19-30.

Blundell TL, Sibanda BL, Montalvão RW, Brewerton S, Chelliah V, Worth CL, Harmer NJ, Davies O, and Burke D: **Structural biology and bioinformatics in drug design: opportunities and challenges for target identification and lead discovery**. *Philos T Roy Soc B* 2006, **361**:413-423.

Boeckmann B, Blatter MC, Famiglietti L, Hinz U, Lane L, Roechert B, Bairoch A: **Protein variety and functional diversity: Swiss-Prot annotation in its biological context**. *C R Biol* 2005, **328:**882-899.

Braconi QS, Orchard S: **The annotation of both human and mouse kinomes in UniProtKB/Swiss-Prot: one small step in manual annotation, one giant leap for full comprehension of genomes**. *Mol Cell Proteomics* 2008, **7:**1409-1419.

Cabrera-Martinez R, Mason JM, Setlow B, Waites WM, Setlow P: **Purification and amino acid sequence of two small, acid-soluble proteins from *Clostridium bifermentans* spores.** *FEMS Microbiol Lett* 1989, **61**:139-144.

Cabrera-Martinez RM, Setlow P: **Cloning and nucleotide sequence of three genes coding for small, acid-soluble proteins of *Clostridium perfringens* spores.** *FEMS Microbiol Lett* 1991, **77**:127-132.

Ceol A, Chatr-Aryamontri A, Santonico E, Sacco R, Castagnoli L, Cesareni G: **DOMINO: a database of domain-peptide interactions**. *Nucleic Acids Res* 2007, **35:**D557-D560.

Charles L, Carbone I, Davies KG, Bird D, Burke M, Kerry BR, Opperman CH: **Phylogenetic analysis of *Pasteuria penetrans* by use of multiple genetic loci.** *J Bacteriol* 2005, **187**:5700-5708.

Cohen AM, Hersh WR: **A survey of current work in biomedical text mining**. *Brief Bioinform* 2005, **6:**57-71.

Conzelmann H, Gilles E: **Dynamic pathway modeling of signal transduction networks: a domain-oriented approach**. *Method Mol Biol* 2008, **484**:559-578.

Couture JF, Collazo E, Hauk G, Trievel RC: **Structural basis for the methylation site specificity of SET7/9**. *Nat Struct Mol Biol* 2006, **13:**140-146.

Dayhoff M: **Atlas of Protein Sequence and Structure,** 1965.

Darland G, Brock TD: ***Bacillus acidocaldarius* sp. nov., an acidophilic thermophilic spore-forming bacterium.** *J Gen Microbiol* 1971, **67**:9-15.

Descorps-Declère S, Barba M, Labedan B: **Matching curated genome databases: a non trivial task**. *BMC Genomics* 2008, **9**:501.

Diella F, Cameron S, Gemund C, Linding R, Via A, Kuster B, Sicheritz-Ponten T, Blom N, Gibson TJ: **Phospho.ELM: A database of experimentally verified phosphorylation sites in eukaryotic proteins**. *BMC Bioinformatics* 2004, **5**:79.

Diella F, Chabanis S, Luck K, Chica C, Ramu C, Nerlov C, Gibson TJ: **KEPE--a motif frequently superimposed on sumoylation sites in metazoan chromatin proteins and transcription factors.** *Bioinformatics* 2009, **25**:1-5.

Diella F, Gould CM, Chica C, Via A, Gibson TJ: **Phospho.ELM: a database of phosphorylation sites - update 2008**. *Nucleic Acids Res* 2008, **36:**D240-D244.

Donahue JP, Vetter ML, Mukhtar NA, D'Aquila RT: **The HIV-1 Vif PPLP motif is necessary for human APOBEC3G binding and degradation**. *Virology* 2008, **377:**49-53.

Doolittle RF: **Similar amino acid sequences: chance or common ancestry?** *Science* 1981, **214**:149-159.

Doolittle RF: **On the trail of protein sequences.** *Bioinformatics 2000,* **16:1 (24-33).**

Doolittle RF: **The Roots of Bioinformatics in Protein Evolution.** *PLoS Computational Biology, 2010,* **6(7).**

Edwards RJ, Davey NE, Shields DC: **CompariMotif: quick and easy comparisons of sequence motifs**. *Bioinformatics* 2008, **24:**1307-1309.

Ekman D, Björklund AK, Frey-Skött J, Elofsson A: **Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions**. *J Mol Biol* 2005, 348:231-243.

Falquet L, Pagni M, Bucher P, Hulo N, Sigrist CJ, Hofmann K, Bairoch A: **The PROSITE database, its status in 2002**. *Nucleic Acids Res* 2002, **30:**235-238.

Fox-Erlich S, Martyn TO, Ellis HJC, Gryk MR: **Delineation and analysis of the conceptual data model implied by the "IUPAC Recommendations for Biochemical Nomenclature."** *Protein Science* 2004, **13:**2559-2563.

Fujii Y, Shimizu T, Toda T, Yanagida M, Hakoshima T: **Structural basis for the diversity of DNA recognition by bZIP transcription factors.** *Nat Struct Biol* 2000, **7**:889-893.

Fukuda M, Moreira JE, Liu V, Sugimori M, Mikoshiba K, Llinas RR: **Role of the conserved WHXL motif in the C terminus of synaptotagmin in synaptic vesicle docking**. *Proc Natl Acad Sci USA* 2000, **97:**14715-14719.

Fulton DL, Li YY, Arenillas DJ, Kwon AT, Wasserman WW: **Improving the specificity of high-throughput ortholog prediction.** *BMC Bioinform* 2006, **7**:270.

Garavelli JS: **The RESID Database of protein structure modifications**. *Nuc Acids Res* 1999, **27:**198-199.

Gattiker A, Hermida L, Liechti R, Xenarios I, Collin O, Rougemont J, Primig M: **MIMAS 3.0 is a multiomics information management and annotation system**. *BMC Bioinformatics* 2009, **10**:151.

Gimona M: **Protein linguistics - a grammar for modular protein assembly?** *Nat Rev Mol Cell Biol* 2006, **7**:68-73.

Goble C, Stevens R: **State of the nation for data integration for bioinformatics.** *J Biomedical Informatics* 2008, **5**:687-693.

Goh CS, Gianoulis TA, Liu Y, Li J, Paccanaro A, Lussier YA, Gerstein M: **Integration of curated databases to identify genotype-phenotype associations**. *BMC Genomics* 2006, **7**:257.

Gong WM, Zhou DH, Ren YL, Wang YJ, Zuo ZX, Shen YP, Xiao FF, Zhu Q, Hong AL, Zhou X, Gao XL, Li TB: **PepCyber: PPEP: a database of human protein-protein interactions mediated by phosphoprotein-binding domains**. *Nucleic Acids Res* 2008, **36**:D679-D683.

Gribskov M, McLachlan AD, Eisenberg D: **Profile analysis: detection of distantly related proteins**. *Proc Natl Acad Sci USA* 1987, **84**:4355-4358.

Gryk MR, Vyas J, Maciejewski MW: **Biomolecular NMR Data Analysis.** *Prog Nucl Magn Reson Spectrosc.* 2010, **56(4)**:329-45.

Guex N, Peitsch MC: **SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling.** *Electrophoresis* 1997, **18**:2714-2723.

Hackett RH, Setlow P: **Properties of spores of *Bacillus subtilis* strains which lack the major small, acid-soluble protein.** *J Bacteriol* 1988, **170**:1403-1404.

Harkiolaki M, Lewitzky M, Gilbert RJC, Jones EY, Bourette RP, Mouchiroud G, Sondermann H, Moarefi I, Feller SM: **Structural basis for SH3 domain-mediated high-affinity binding between Mona/Gads and SLP-76**. *EMBO J* 2003, **22**:2571-2582.

Hasegawa M, Kishino H, Yano T: **Dating of the human-ape splitting by a molecular clock of mitochondrial DNA.** *J Mol Evol* 1985, **22**:160-174.

Hagen: **Naturalists, Molecular Biologists, and the Challenge of Molecular Evolution.** J Hist. Bio. 1999, **32**:323-325.

Henikoff S, Henikoff JG: **Amino acid substitution matrices from protein blocks.** *Proc Natl Acad Sci USA* 1992, **89**:10915-10919.

Holm LF, Sander C: **Protein structure comparison by alignment of distance matrices.** *J Mol Biol* 1993, **233**:123-138.

Hunter L, Cohenw KB: **Biomedical language processing: what's beyond PubMed?** *Mol Cell* 2006, **21**:589-594.

Huo Z, Yang X, Raza W, Huang Q, Xu Y, Sheng Q: **Investigation of factors influencing spore germination of *Paenibacillus polymyxa* ACCC10252 and SQR-21.** *Appl Microbiol Biotechnol* 2010, **87**:527-536.

Goodman L: **Hypothesis-limited research** [editorial]. *Genome Res* 1999, **9**:673-674.

Hermann T, Guntert P, Wuthrich K: **Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithms DYANA.** *J Mol. Bio,* **319(1):**209-207.

IUPAC-IUB Commission on Biochemical Nomenclature (CBN): **Abbreviations and symbols for nucleic acids, polynucleotides and their constituents. Recommendations 1970**. *Biochem J* 1970, **120:**449-454.

Jia CYH, Nie J, Wu CG, Li CJ, Li SSC: **Novel Src homology 3 domain-binding motifs identified from proteomic screen of a pro-rich region**. *Mol Cell Proteomics* 2005, **4:**1155-1166.

Kaga C, Okochi M, Tomita Y, Kato R, Honda H: **Computationally assisted screening and design of cell-interactive peptides by a cell-based assay using peptide arrays and a fuzzy neural network algorithm**. *Biotechniques* 2008, **44:**393-402.

Kaneko T, Li L, Li SS: **The SH3 domain--a family of versatile peptide--and protein-recognition module**. *Front Biosci* 2008, **13**:4938-4952.

Kaushansky A, Gordus A, Chang B, Rush J, MacBeath G: **A quantitative study of the recruitment potential of all intracellular tyrosine residues on EGFR, FGFR1 and IGF1R**. *Mol Biosyst* 2008, **4:**643-653.

Kawaji H, Hayashizaki Y: **Genome annotation**. *Methods Mol Biol* 2008, **452:**125-139.

Kitano H: **Systems biology: a brief overview**. *Science* 2002, **295**:1662-1664.

Knight RD, Freeland SJ, Landweber LF: **Selection, history and chemistry: the three faces of the genetic code.** *Trends Biochem Sci. 1999.* **24(6)**:241-7

Kochiwa H, Tomita M, Kanai A: **Evolution of ribonuclease H genes in prokaryotes to avoid inheritance of redundant genes.** *BMC Evol Biol* 2007, **7**:128.

Kopp J, Schwede T: **The SWISS-MODEL repository of annotated three-dimensional protein structure homology models.** *Nucleic Acids Res* 2004, **32**:D230-D234.

Koradi R, Billeter M, Wuthrich K: **MOLMOL: a program for display and analysis of macromolecular structures**. *J Mol Graph* 1996, **14:**51-55.

Labarga A, Valentin F, Anderson M, Lopez R: **Web services at the European bioinformatics institute.** *Nucleic Acids Res* 2007, **35**:W6-11.

Landau M, Mayrose I, Rosenberg Y, Glaser F, Martz E, Pupko T, Ben Tal N: **ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures.** *Nucleic Acids Res* 2005, **33**:W299-W302.

Lederberg J: **The Transformation of Genetics by DNA: An Anniversary Celebration of Avery, Macleod and Mccarty (1944)** *Genetics. 1994* **136**(2): 423–426.

Lee KS, Bumbaca D, Kosman J, Setlow P, Jedrzejas MJ: **Structure of a protein-DNA complex essential for DNA protection in spores of *Bacillus* species.** *Proc Natl Acad Sci USA,* 2008 **105**:2806-2811.

Lesk A.M: **Computational and Molecular Biology:** *Encyclopedia of Computer Science and Technology, 1994 **31**:101-165.*

Li SS: **Specificity and versatility of SH3 and other proline-recognition domains: structural basis and implications for cellular signal transduction**. *Biochem J* 2005, **390:**641-653.

Lindwasser OW, Smith WJ, Chaudhuri R, Yang P, Hurley JH, Bonifacino JS: **A diacidic motif in human immunodeficiency virus type 1 Nef is a novel determinant of binding to AP-2**. *J Virol* 2008, **82**:1166-1174.

Lipton P: **Testing hypotheses: prediction and prejudice**. *Science* 2005, 307:219-221.

Liu Q, Berry D, Nash P, Pawson T, McGlade CJ, Li SSC: **Structural basis for specific binding of the gads SH3 domain to an RxxK motif-containing SLP-76 peptide: a novel mode of peptide recognition**. *Mol Cell* 2003, **11:**471-481.

Loshon CA, Beary KE, Gouveia K, Grey EZ, Santiago-Lara LM, Setlow P: **Nucleotide sequence of the *sspE* genes coding for g-type small, acid-soluble spore proteins from the round-spore-forming bacteria *Bacillus aminovorans*, *Sporosarcina halophila* and *S. ureae*.** *Biochem Biophys Acta* 1998, **1396**:148-152.

Loshon CA, Fliss ER, Setlow B, Foerster HF, Setlow P: **Cloning and nucleotide sequencing of genes for small, acid-soluble spore proteins of *Bacillus cereus*, *Bacillus stearothermophilus*, and "*Thermoactinomyces thalpophilus*".** *J Bacteriol* 1986, **167**:168-173.

Magill NG, Loshon CA, Setlow P: **Small, acid-soluble, spore proteins and their genes from two species of *Sporosarcina*.** *FEMS Microbiol Lett* 1990, **72**:293-298.

Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S: **The protein kinase complement of the human genome**. *Science* 2002, **298**:1912-1934.

McCarty, Maclyn, and Oswald T. Avery. 1946. STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES. *The Journal of Experimental Medicine* 83, no. 2 (January 31): 97-104.

Marchler-Bauer A, Anderson JB, DeWeese-Scott C, Fedorova ND, Geer LY, He S, Hurwitz DI, Jackson JD, Jacobs AR, Lanczycki CJ, Liebert CA, Liu C, Madej T, Marchler GH, Mazumder R, Nikolskaya AN, Panchenko AR, Rao BS, Shoemaker BA, Simonyan V, Song JS, Thiessen PA, Vasudevan S, Wang Y, Yamashita RA, Yin JJ, Bryant SH: **CDD: a curated Entrez database of conserved domain alignments**. *Nuc Acids Res* 2003, **31**:383-387.

Mauchline TH, Mohan S, Davies KG, Schaff JE, Opperman CH, Kerry BR, Hirsch PR: **A method for release and multiple strand amplification of small quantities of DNA from endospores of the fastidious bacterium *Pasteuria penetrans*.** *Lett Appl Microbiol* 2010, **50**:515-521.

McDonald DM, Chen H, Su H, Marshall BB: **Extracting gene pathway relations using a hybrid grammar: the Arizona Relation Parser**. *Bioinformatics* 2004, 20:3370-3378.

Mihalek I, Res I, Lichtarge O: **A family of evolution-entropy hybrid methods for ranking protein residues by importance.** *J Mol Biol* 2004, **336**:1265-1282.

Moore AD, Björklund AK, Ekman D, Bornberg-Bauer E, Elofsson A.: **Arrangements in the modular evolution of proteins.** *Trends Biochem Sci. 2008, **33**:444-51.*

Montelione G, Arrowsmith C, Girvin M, Kennedy M, Markley J, Powers R, Prestegard J, Syperski T: **Unique opportunites for NMR methods in structural genomics.** *J Struct Funct Genomics* 2009, **10(2)**:101-106.

Morgan DH, Kristensen DM, Mittelman D, Lichtarge O: **ET viewer: an application for predicting and visualizing functional sites in protein structures.** *Bioinformatics* 2006, **22**:2049-2050.

Morton CJ, Pugh DJ, Brown EL, Kahmann JD, Renzoni DA, Campbell ID: **Solution structure and peptide binding of the SH3 domain from human Fyn**. *Structure* 1996, **4:**705-714.

Muirhead H, Perutz M: **Structure of hemoglobin. A three-dimensional fourier synthesis of reduced human hemoglobin at 5.5 Å resolution**. *Nature* 1963. **199** (4894): 633–38.

Muller HM, Kenny EE, Sternberg PW: **Textpresso: an ontology-based information retrieval and extraction system for biological literature**. *PLoS Biol* 2004, **2:**e309.

Muslin AJ, Tanner JW, Allen PM, Shaw AS: **Interaction of 14-3-3 with signaling proteins is mediated by the recognition of phosphoserine**. *Cell* 1996, **84:**889-897.

Neduva V, Russell RB: **DILIMOT: discovery of linear motifs in proteins**. *Nucleic Acids Res* 2006, **34:**W350-W355.

Nicholson WL, Setlow P: **Sporulation, germination and outgrowth.** In: Harwood CR, Cutting SM, eds. *Molecular biological methods for Bacillus.* Chichester, United Kingdom: Wiley, 1990:391-450.

Nicholson WL, Sun D, Setlow B, Setlow P: **Promoter specificity of s^G-containing RNA polymerase from sporulating cells of *Bacillus subtilis*: identification of a group of forespore-specific promoters.** *J Bacteriol* 1989, **171**:2708-2718.

Nirenberg, M, P Leder, M Bernfield, R Brimacombe, J Trupin, F Rottman, and C O'Neal. 1965. RNA codewords and protein synthesis, VII. **On the general nature of the RNA code.** *Proceedings of the National Academy of Sciences of the United States of America* 53, no. 5 (May): 1161-1168.

Ronald J. Nowling, Jay Vyas, Gerard Weatherby, Matthew W. Fenwick, Heidi J.C. Ellis, Michael R. Gryk: **CONNJUR spectrum translator: an open source application for reformatting NMR spectral data.** *JBio NMR.* 2011 May, 50(1):83-89.

Obenauer JC, Cantley LC, Yaffe MB: **Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs**. *Nucleic Acids Res* 2003, **31**:3635-3641.

Olsen SK, Li JYH, Bromleigh C, Eliseenkova AV, Ibrahimi OA, Lao ZM, Zhang FM, Linhardt RJ, Joyner AL, Mohammadi M: **Structural basis by which alternative splicing modulates the organizer activity of FGF8 in the brain.** *Gen Dev* 2006, **20**:185-198.

Paredes-Sabja D, Setlow P, Sarker MR: **Germination of spores of *Bacillales* and *Clostridiales* species: mechanisms and proteins involved.** *Trends Microbiol* in press.

Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi TKB, Gronborg M, Ibarrola N, Deshpande N, Shanker K, Shivashankar HN, Rashmi BP, Ramya MA, Zhao ZX, Chandrika KN, Padma N, Harsha HC, Yatish AJ, Kavitha MP, Menezes M, Choudhury DR, Suresh S, Ghosh N, Saravana R, Chandran S, Krishna S, Joy M, Anand SK, Madavan V, Joseph A, Wong GW, Schiemann WP, Constantinescu SN, Huang LL, Khosravi-Far R, Steen H, Tewari M, Ghaffari S, Blobe GC, Dang CV, Garcia JGN, Pevsner J, Jensen ON, Roepstorff P, Deshpande KS, Chinnaiyan AM, Hamosh A, Chakravarti A, Pandey A: **Development of human protein reference database as an initial platform for approaching systems biology in humans**. *Genome Res* 2003, **13**:2363-2371.

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE: **UCSF Chimera--a visualization system for exploratory research and analysis.** *J Comput Chem* 2004, **25**:1605-1612.

Pieper U, Eswar N, Braberg H, Madhusudhan MS, Davis FP, Stuart AC, Mirkovic N, Rossi A, Marti-Renom MA, Fiser A, et al.: **MODBASE, a database of annotated comparative protein structure models, and associated resources.** *Nucleic Acids Res* 2004, **32**:D217-D222.

Pires JR, Hong X, Brockmann C, Volkmer-Engert R, Schneider-Mergener J, Oschkinat H, Erdmann R: **The ScPex13p SH3 domain exposes two distinct binding sites for Pex5p and Pex14p**. *J Mol Biol* 2003, **326**:1427-1435.

Pruitt KD, Maglott DR: **RefSeq and LocusLink: NCBI gene-centered resources**. *Nucleic Acids Res* 2001, **29**:137-140.

Pruitt K, Tatiana T, Maglott D: **NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucl. Acids Res. (2005),* 33(suppl 1): D501-D504

Puntervoll P, Linding R, Gemund C, Chabanis-Davidson S, Mattingsdal M, Cameron S, Martin DMA, Ausiello G, Brannetti B, Costantini A, Ferre F, Maselli V, Via A, Cesareni G, Diella F, Superti-Furga G, Wyrwicz L, Ramu C, McGuigan C, Gudavalli R, Letunic I, Bork P, Rychlewski L, Kuster B, Helmer-Citterich M, Hunter WN, Aasland R, Gibson TJ: **ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins**. *Nucleic Acids Res* 2003, **31:**3625-3630.

Quirk PG: **A gene encoding a small, acid-soluble spore protein from alkalaphilic *Bacillus firmus* OF4.** *Gene* 1993, **125**:81-83.

Rajasekaran S, Balla S, Gradie P, Gryk MR, Kadaveru K, Kundeti V, Maciejewski MW, Mi T, Rubino N, Vyas J, Schiller MR: **Minimotif miner 2nd release: a database and web system for motif search**. *Nucleic Acids Res* 2009, **37:**D185-D190.

Raman S, Oliver L, Rossi P, 5. Xu Wang, James Aramini, Gaohua Liu, Theresa A. Ramelot, Alexander Eletsky, Thomas Szyperski, Michael A. Kennedy, James Prestegard, Gaetano T. Montelione, David Baker: **NMR Structure Determination for Larger Proteins Using Backbone-Only Data1.** *Science 2010*:**327**:1014-1018.

Ramirez-Flandes S, Ulloa O: **Bosque: integrated phylogenetic analysis software.** *Bioinfor* 2008, **24**:2539-2541.

Rawal N, Rajpurohit R, Lischwe MA, Williams KR, Paik WK, Kim S: **Structural specificity of substrate for S-adenosylmethionine protein arginine N-methyl transferases.** *Biochim Biophys Acta* 1995, **1248:**11-18.

Rawlings ND, Morton FR, Barrett AJ: **MEROPS: the peptidase database**. *Nucleic Acids Res* 2006, **34:**D270-D272.

Reed JL, Famili I, Thiele I, Palsson BO: **Towards multidimensional genome annotation**. *Nature Rev Genet* 2006, **7:**130-141.

Reeves GA, Talavera D, Thornton JM: **Genome and proteome annotation: organization, interpretation and integration**. *J R Soc Interface* 2009, **6:**129-147.

Reichman C, Singh K, Liu Y, Singh S, Li H, Fajardo JE, Fiser A, Birge RB: **Transactivation of Abl by the Crk II adapter protein requires a PNAY sequence in the Crk C-terminal SH3 domain**. *Oncogene* 2005, **24:**8187-8199.

Rhoads AR, Friedberg F: **Sequence motifs for calmodulin recognition**. *FASEB* 1997, **11**:331-340.

Richmand WK: **The Educational Industry.** *Methuen 1969.*

Rockwell NC, Lagarias JC: **Flexible mapping of homology onto structure with homolmapper.** *BMC Bioinformatics* 2007, **8**:1-13.

Sankoff, D: **Matching sequences under deletion/insertion constraints**. *Proceedings of the National Academy of Sciences of the USA*. **1**: 4–6.

Sayers EW, Barret T, Benson DA, et al.: **Database resources of the National Center for Biotechnology Information.** *Nucl Acids Res* 2010, **38**:D5-16.

Sargeant D, Deverasetty S1, Luo Y, Baleta AV1, Zobrist S, Rathnayake V, Russo JC, Vyas J, Muesing MA, and Schiller MR : **HIVToolbox, an integrated web application for investigating HIV.** *Plos One*

Schiller MR, Chakrabarti K, King GF, Schiller NI, Eipper BA, Maciejewski MW: **Regulation of RhoGEF activity by intramolecular and intermolecular-SH3 interactions.** *J Biol Chem* 2006, **281:**17774-17786.

Schneider TD, Stephens RM: **Sequence logos: a new way to display consensus sequences**. *Nucleic Acids Res* 1990, **18:**6097-6100.

Schwede T, Kopp J, Guex N, Peitsch MC: **SWISS-MODEL: an automated protein homology-modeling server.** *Nucleic Acids Res* 2003, **31**:3381-3385.

Yang S, Lange O, Delaglio F: **Consistent blind protein structure generation from NMR chemical shift data.** *Proc Natl Acad Sci USA* 2008. **105**:4685-4690

Shuch P: **The Only Game In Town**. Journal of Future Studies, February 2004, 8(3):55-60.

Setlow B, Cowan AE, Setlow P: **Germination of spores of *Bacillus subtilis* with dodecylamine.** *J Appl Microbiol* 2003, **95**:637-648.

Setlow P: **I will survive: DNA protection in bacterial spores.** *Trends Microbiol* 2007, **15**:172-180.

Setlow P: **Purification and characterization of additional low-molecular weight basic proteins degraded during germination of *Bacillus megaterium* spores.** *J Bacteriol* 1978, **136**:331-340.

Setlow P: **Small acid-soluble, spore proteins of *Bacillus* species: structure, synthesis, genetics, function and degradation.** *Ann Rev Microbiol* 1988, **42**:319-338.

Setlow P: **Spores of *Bacillus subtilis*: their resistance to radiation, heat and chemicals.** *J Appl Microbiol* 2006, **101**:514-525.

Sherman BT, Huang dW, Tan Q, Guo Y, Bour S, Liu D, Stephens R, Baseler MW, Lane HC, Lempicki RA: **DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis**. *BMC Bioinformatics* 2007, **8:**426.

Shida O, Takagi H, Kawadaki K, Komagata K: **Proposal for two new genera, *Brevibacillus* gen. nov. and *Aneurinibacillus* gen. nov.** *Int J System Bacteriol* 1996, **46**:939-946.

Sidhu A, Dillon TS, Sidhu BS, Chang E: **Protein ontology; seamless protein data integration**. *Mol Cell Proteomics* 2005, **4:**S84.

Smith, Temple F.; and Waterman, Michael S.: **Identification of Common Molecular Subsequences**. *Journal of Molecular Biology.* **147**:195–197.

Stein LD: **Integrating biological databases.** *Nature Rev. Genet.* 2003. 4(5):337-45.

Simsion G, Witt G: **Data Modeling Essentials**, The Morgan Kaufmann Series in Data Management Systems, REV, 3, Book, ISBN: 0126445516

Songyang Z: **Recognition and regulation of primary-sequence motifs by signaling modular domains**. *Prog Biophys Mol Biol* 1999, **71:**359-372.

Songyang Z, Shoelson SE, Mcglade J, Olivier P, Pawson T, Bustelo XR, Barbacid M, Sabe H, Hanafusa H, Yi T, Ren R, Baltimore D, Ratnofsky S, Feldman RA, Cantley LC: **Specific motifs recognized by the Sh2 domains of Csk, 3Bp2, Fps Fes, Grb-2, Hcp, Shc, Syk, and Vav**. *Mol Cell Biol* 1994, **14:**2777-2785.

Stark GR, WR Taylor: **Control of the G2/M transition**. *Mol Biotechnol* 2006, **32**: 227-248.

Stevens R, Goble C, Baker P, Brass A. **A classification of tasks in bioinformatics.** *Bioinformatics* 2001, 17:180-188.

Strasser, BJ: **Collecting, comparing, and computing sequences: the making of Margaret O. Dayhoff's Atlas of Protein Sequence and Structure, 1954-1965.** J Hist Biol 43(4):623-60

Stravinsky, Igor: **Poetics of Music in the Form of Six Lessons.** Cambridge, MA: Harvard University Press. *1947* OCLC 155726113.

Strömbergsson H, GJ Kleywegt: **A chemogenomics view on protein-ligand spaces**. *BMC Bioinformatics* 2009, 10(suppl 6):S13.

Taylor Ronald C: **An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics.** *BMC Bioinformatics 2010,* 11:12.

Sun D, Setlow P: **Cloning and nucleotide sequencing of genes for the second type of small, acid-soluble spore proteins of *Bacillus cereus*, *Bacillus stearothermophilus*, and "*Thermoactinomyces thalpophilus*".** *J Bacteriol* 1987, **169**:3088-3093.

Ulrich E, Akutsu H, Doreleijers J, Harano Y, Ioannidis Y, Lin J, Livny M, Mading S, Maziuk D, Miller Z, Nakatani E, Schulte C, Tolmie D, Wenger K,, Yao H, Markley J: **BioMagResBank.** *Nucleic Acids Research* (2007), 36, D402-D408

Venkatesh TV, Harlow HB: **Integromics: challenges in data integration.** *Genome Biol* [serial online]. 2002, **3**:reports4027-4027.3. Available from http://genomebiology.com/2002/3/8/reports/4027. Accessed August 16, 2011.

Venn J: **On the diagrammatic and mechanical representation of propositions and reasonings.** *J Science* 1880, **9**:1-18.

Ventner C Adams EW, Myers PW, Li RJ, Mural GG, Sutton, HO, Smith, et al.: **The Sequence of the human genome.** *Science* 2001, **291**:1304-1351.

Vyas J, *Personal Communication with NCBI Administrators,* Jan 22, 2008.

Vyas J, *Personal Communication with UCSC Genome Database Curators,* Apr 7, 2008.

Vyas J, Nowling RJ, Maciejewski MW, Rajasekaran S, Gryk MR, Schiller MR: **A proposed syntax for Minimotif Semantics, version 1**. *BMC Genomics* 2009, **10**:360.

Vyas J, Gryk MR, Schiller MR: **Venn, a tool for titrating sequence conservation onto protein structures.** *Nucleic Acids Res.* 2009; **(18)**:e124.

Vyas J, RJ Nowling email, Thomas Meusburger2 email, David Sargeant2 email, Krishna Kadaveru1 email, Michael R Gryk1 email, Vamsi Kundeti3 email, Sanguthevar Rajasekaran3 email and Martin R Schiller1,2 emailMR: **A proposed syntax for Minimotif Semantics, version 1**. *BMC Genomics* 2009, **10**:360.

Watson JD, Crick, FH. **Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid.** *Nature* 1953, **171:** 737-738.

Wheeler DL, Chappey C, Lash AE, Leipe DD, Madden TL, Schuler GD, Tatusova TA, Rapp BA: **Database resources of the National Center for Biotechnology Information**. *Nucleic Acids Res* 2000, **28:**10-14.

Wu HW, Maciejewski MW, Takebe S, King SM: **Solution structure of the Tctex1 dimer reveals a mechanism for dynein-cargo interactions.** *Structure* 2005, **13**:213-223.

Yoon J-H, Kim I-G, Shin Y-K, Park Y-H: **Proposal of the genus *Thermoactinomyces sensu stricto* and three new genera, *Laceyalla*, *Thermoflavimicrobium* and *Seinonella*, on the basis of phenotypic, phylogenetic and chemotaxonomic analyses.** *Int J Syst Evol Microbiol* 2005, **55**:395-400.

Yoon J-H, Park YH: **Phylogenetic analysis of the genus *Thermoactinomyces* based on 16S rDNA sequences.** *Int J System Evol Microbiol* 2000, **50**:1081-1086.

Yuan K, Johnson WC, Tipper DJ, Setlow P: **Comparison of various properties of low-molecular weight proteins from dormant spores of various *Bacillus* species.** *J Bacteriol* 1981, **146**:965-971.

Appendix A. Additional Material for Chapter 1

Table A1. Physical to conceptual data model mapping

Physical Model	Conceptual Model	Purpose
Motif	Motif	Defines the motif sequence, and any post-translational modification
ref_knownactivity,motif_source	Activity	Defines the motifs activity for a given annotated motif.
ref_molecule, motif_source	Target	Defines the biological target of an annotated motif.
ref_homologene_2_gene_protein	RefSeq	Defines the RefSeq record for a motif containing protein or its target.
ref_homologene_2	HomoloGene	Defines a HomoloGene cluster for any protein.
ref_domain	CDD	Defines types of protein domains

Table A2 defines the rules for generating human readable annotations from the structured attributes of the minimotif syntax. The syntactical attributes can be acquired by joining tables in the database. The value of different attributes in each condition for a minimotif determines which rule is used.

SH3 Binding Motif Clustering

In order to determine SH3 domain binding motifs, a query against the ref_knownactivity, ref_molecule, motif_source, motif and ref_domain tables was executed to join their data (*query 1)*. The resultant cluster of motif *sequences* from this data set consisted of 741 distinct sequences (69 consensus sequences and 672 instances), with 59 *Target* (SH3 containing) proteins and 372 source (*Motif* containing) proteins. At this point, we have utilized our semantic model of

Table A2. Rules used to regenerate annotations from database tables.

Rule #	Condition	Rule
1	([Activity] = binds) AND ([Required Modification] = Instance) AND (Target Name domain = empty or null) AND ([Required Modification] does not = none)	[Motif Sequence] in [Motif source name] [Activity] [Target Name]; [Required Modification]
2	([Activity] = binds) AND ([Required Modification] = Instance) AND (Target Name domain = empty or null) AND ([Required Modification] = none)	[Motif Sequence] in [Motif source name] [Activity] [Target Name]
3	([Activity] = binds) AND ([Required Modification] = Instance) AND (Target Name domain = is not empty or null) AND ([Target domain position] = empty or null)	[Motif Sequence] in [Motif source name] [Activity] the Target Name domain Target Name domain of [Target Name]; [Required Modification]
4	([Activity] = binds) AND ([Required Modification] = Instance) AND (Target Name domain = is not empty or null) AND ([Target domain position] = is not empty or null)	[Motif Sequence] in [Motif source name] [Activity] the [Target domain position] Target Name domain Target Name domain of [Target Name]; [Required Modification]
5	([Activity] = binds) AND ([Subactivity] contains trafficked) AND ([Required Modification] = Instance) AND ([Required Modification] = none)	[Motif Sequence] in [Motif source name] binds [Target Name] and is [Subactivity] [Subcellular Localization]
6	([Activity] = binds) AND ([Subactivity] contains trafficked) AND ([Required Modification] = Instance) AND ([Required Modification] is not = none)	[Motif Sequence] in [Motif source name] binds [Target Name] and is [Subactivity] [Subcellular Localization]; [Required Modification]

Table A2 (continued)

Rule #	Condition	Rule
7	([Activity] = requires) AND ([Required Modification] = Instance) AND ([Required Modification] = none)	[Motif Sequence] [Subactivity] requires [Required Modification] motif in [Motif source name]; Target Name is [Target Name]
8	([Activity] = requires) AND ([Required Modification] = Instance) AND ([Required Modification] does not = none)	[Motif Sequence] [Subactivity] requires [Required Modification] motif in [Motif source name]; Target Name is [Target Name]; [Required Modification]
9	([Activity] = modifies) AND ([Required Modification] = Instance) AND ([Required Modification] = none)	[Motif Sequence] in [Motif source name] is [Subactivity] by [Target Name]; [Activity Modification]
10	([Activity] = binds) AND ([Required Modification] = Instance) AND ([Required Modification] does not = none)	[Motif Sequence] in [Motif source name] is [Subactivity] by [Target Name]; [Activity Modification]; [Required Modification]

minimotif function to derive a data set resulting from a very specific linguistic analysis which can now be analyzed for minimotif groupings. Several database procedures were needed for this analysis (queries 1-9).

Initially, consensus motifs were separated from motif instances using query 1. This statement returned a series of sequence instances in MnM 2 which bind the SH3 domain of a *Target* protein, along with the name of that *Target* protein, e.g.

AKLKPGAPLRPKLN	ABL
AKLKPGAPVRSKQL	Grb2
AKPKKAPKSPAKA	Nck1

Table A3. Queries for SH3 binding minimotif analysis

Query number	Syntax
1	*Select sequence, '#', ref_molecule.name from motif, motif_source, ref_knownactivity, ref_molecule, ref_domain where motif_source.motif=motif.id, ref_molecule.id=motif_source.target, and ref_molecule.refDomain = ref_domain.id and ref_domain.domain = 'SH3' and ref_knownactivity.Activity='binds' and motif.type IS NOT 'Consensus'*
2	*Select motifClass,count(*),(select count(*) from motif_comparison), avg(score) from motif_comparison where score > 1 group by motifClass order by count(*)*
3	*Select sh3_group.rxp, count(0)/(select count(*) from lexica) from sh3_group join lexica where lexica.sequence regexp (sh3_group.rxp) group by sh2_group.rxp union select 'NOT PXXP', count(0), count(0)/(select count(*) from lexica) from lexica where not isPxxP(lexica.sequence)*
4	*declare totalresidues int;select sum(length(m.sequence)) into totalresidues from sh3_binding_motifs_sandbox m; select a.letter,sum(substrCount(s.sequence,a.letter)) rawTotalCount, 100*sum(substrCount(s.sequence,a.letter))/totalresidues as percentComposition, 100*sum(substrCount(s.sequence,a.letter)>0)/totalresidues as rawAmountContaining, (100*sum(substrCount(s.sequence,a.letter))/totalresidues)/enric.perc ent as percentCompositionNormalizedToProteome from sh3_binding_motifs_sandbox s, ref_amino_acid a, ref_aa_enrichment_human_proteome enrich where enric.aa=a.`letter`group by a.letter*
5	***q**Select motifClass,count(*),(select count(*) from motif_comparison), avg(score) from motif_comparison where score > 1 group by motifClass order by count(*)*
6	***number of SH3 containing proteins in human proteome:** 'Select distinct ref_homologene_2 from ref_homologene_2 h, ref_homologene_2_gene_domain d,ref_homologene_2_gene g where domain =<domain> and d.ref_homologene_gene=g.id and g.ref_homologene_2=h.id'*

Table A3 (continued)

Query number	Syntax
7	***number of unique SH3 binding sequences:*** *'Select distinct sequence from motif,motif_source,ref_molecule,ref_knownactivity a where a.Activity ='binds' and motif_source.knownActivity=a.id and motif_source.motif=motif.id and ref_molecule.id=motif_source.target and ref_molecule.ref_domain=(select id from ref_domain where domain='SH3') order by sequence and not sequence regexp('x') and not sequence like '%[%''*
8	***charged character of SH3 binding sequences*** *: 'Select avg(getPeptideCharge(s.sequence)) from human_proteome as s UNION select avg(getPeptideCharge(s.sequence)) from distinct sh3_binding_lexica_type group by s.sequence regexp ('[KR]..[KR]')'*
9	*'Select avg(s.cnt) from (select count(*) as cnt from motif_source_motif_group where group_title='SH3' and motif regexp(group_rxp) group by motif) s'*

By running query 1 again, this time omitting the final 'and' clause, we extract minimotif consensus sequences, where the purpose of the '#' is to format the data on export so that it is directly compatible with the Comparimotif program which was used for comparing instances against consensa (Edwards et al., 2008).

By utilizing the Comparimotif program to compare minimotif instance data against consensa, and integrating this data set to MnM, we could now cross-query between the results of a global Comparimotif analysis of the motifs using query 2. This revealed the most common SH3 binding motif consensa. This analysis revealed a variety of such relationships between consensus

sequences and instances. We ranked relationships by using Comparimotif's Shannon's Information Content based score with a cutoff value of 2.0 since low scores did not show meaningful relationships between consensus sequences and instances (Edwards et al., 2008). Considering only scores above this cutoff, we then tabulated a relevance score for important consensus sequences (Table A4). We define 'Relevance Percent' as the ratio of the number of Comparimotif calculated matches for a consensus by the total amount of distinct instances variants in our database for SH3 binding peptides. For example, a consensus sequence which matched to every SH3 binding instance sequence in MnM 2 would have a score of 100%.

Table A4. Frequencies of exact matching instances / consensus sequences in database.

Consensus	Number
KKPP	7
PxxxPR	183
PxxDY	2
PxxP	1305
PxxPx[KR]	972
RxxPxxP	308
RKxxYxxY	3
WxxFxLE	1
[HKR]xxHKR]	495
KPTVY	2

Table A5 indicates importance of all the consensus sequences in the minimotif database in terms of their frequency. The PxxP motif, for example, was an important class since it had the highest frequency. The second most important matches, PxxPx[KR] and PxxPxK are known class II SH3 binding motifs.

Table A5. Consensus sequence relevance ranking.

Consensus	Relevance Percent
Px[IV]PPR	3.0
PLPxLP	3.8
[KR]xxxxKx[KR][KR]	3.8
PxPPxRxSSL	4.6
RxLPxLP	4.6
PxPPxRxxSL	5.2
RxxK	7.9
KxxK	8.7
Px[AP]x[PV]R	22.1
PxLPxK	12.6
[KR]xLPxxP	18.8
PxxxPR	20.7
RxxPxxxP	24.9
Px[AP]xxR	33.6
PxxPxK	35.0
PxxPx[KR]	74.0
PxxP	89.1

Many of the consensus sequences were related as are the two class II motifs above. Therefore we used Cytoscape to visualize all consensus sequences related to instances and grouped motifs that had common sets of instances (Shannon et al., 2003). The visualization of matches using Cytoscape

allowed us to identify several important consensus sequences. Although the implementation of Shannon Information Content scoring gives us a valuable initial screen of motif significance, we also used regular expression matching in SQL to identify "exact" matches. Since this was an important query for our analysis, we embedded it in our database as a view (a table with all contents dynamically derived from other tables).

This analysis resulted in ten different consensus groups (PxxDY, PxxP, [HKR]xx[HKR], PxxxPR, PxxPx[KR], RxxPxxP, WxxxFxxLE, RKxxYxxY, KKPP, and KPTVY). The results from query 3 identified PxxP, RxxPxxP and PxxPx[KR] as the most common motifs (Table 4). However, PxxxPR, BxxB, and [HKR]xx[HKR] may also be highly significant SH3 binding motifs that bind to distinct sites. Additionally, KKPP, WxxFxxLE, PxxDY, and RKxxYxxY are underrepresented in our database and their broader significance in binding SH3 domains will require further study. One limitation with the frequency-based analysis is that the SH3 domains and motifs thus far experimentally examined are biased, as may be the content of our database. We have also evaluated the validity of our motif categorization by comparing the binding sites of different SH3 binding motifs in a structural analysis.

Analysis of Residue Content in SH3 Domain Binding Peptides

Residue content in all SH3 ligands was determined using queries 4 and 5. Query 5 identifies the frequency of each residue in all SH3 binding minimotifs and these numbers were normalized to the frequency of each residue in the human proteome which was identified using query 5. Query 5 stores this data in

a table titled ref_aa_enrichment_human_proteome which has each residue, a percentage value for its enrichment, and its fold enrichment in SH3 binding sequences.

Appendix B: Additional Material for Chapter 3

Identification of Papers with Minimotif Content

In our initial attempts to collect papers from the literature that have minimotif content, we tested several queries. To evaluate each query, a Minimotif Identification Efficiency (MIE) score was calculated. To determine this score, a subset, consisting of 10-20 randomly-selected papers chosen from the results of the search, was selected. MIE is simply the percentage of those papers that have minimotifs. Using MIE and other criteria, a search query is either accepted or rejected. Accepted queries are used to add papers to a paper list in the Minimotif database (see Fig. B1).

In addition to Keyword and MeSH term queries of PubMed, we used several other strategies to identify papers containing minimotif information. These included: author/affiliation searches that identified papers by authors (with their institutional affiliations) of minimotif data-containing papers already in the MnM database, regular expression searches which identified papers with abstracts that contain strings of peptide sequences or consensus sequences using regular expressions, reverse citation searches which identified papers referenced by papers already in the MnM, forward citation searches which identified papers that referenced a paper in the MnM database, journal selection identified which journals have higher probabilities of publishing minimotif papers, and publication year which was used to restrict searches to more recent papers in PubMed. Combined, these strategies were used to build a list of ~130,000 papers that had a MIE score of ~30%.

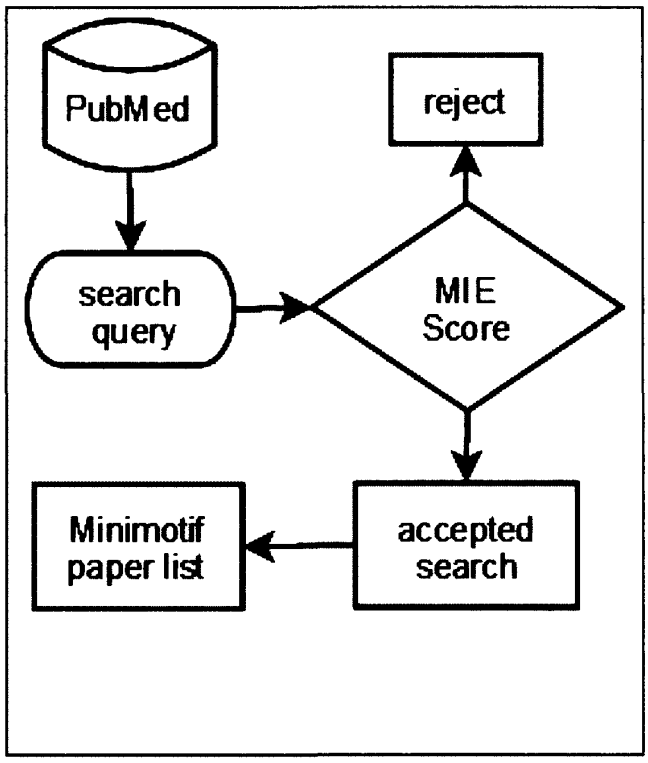


Fig. B1. Strategy for identifying papers with minimotif content.

Automated markup of paper abstracts in MimoSA

Through an integrated database of PubMed abstracts, their lexemes, and several million RefSeq and CDD keywords, MimoSA automates the process of marking up potential key annotation terms which are key indicators of minimotif meta data in abstracts [1, 2]. Automatically detectable elements of a minimotif annotation include activity terms, minimotif interaction domains, minimotif target or source proteins, and minimotif sequence information.

Minimotif detection. In order to detect terms that might contain minimotif sequence or consensus residue information, we derived a regular expression for amino acid sequences. To speed up the process of minimotif sequence detection

by users, papers are automatically screened, and all text sequences that conform to the following sequence by regular expression are highlighted and flagged.

(([Xx])|(Gly)|(Ala)|(Val)|(Leu)|(Ile)|(Met)|(Phe)|(Trp)|(Pro)|(Ser)|(Thr)|(Cys)|(Tyr)|(Asn)|(Gln)|(Lys)|(Arg)|(His)|(Asp)|(Glu)|(Lys)|(Thr)|(Trp)|\p{Punct}|)-?){3,15}.

Activity detection. To speed up the process of activity annotation, key terms for suggested minimotif activities in a paper are automatically highlighted. These terms come from the several hundred discrete sub-activity term definitions in the MnM database. In addition, the words "binds", "modifies", and "required" are highlighted.

Interaction partners and targets. In order to detect important domains and / or proteins, a string-matching algorithm that searches for words which are associated with gene names, aliases, or RefSeq protein names is applied to all abstracts. Domains and proteins are highlighted in different colors. This was useful for annotation as many targets of minimotifs are proteins and more specifically domains within proteins.

Pseudocode for Paper Scoring Algorithm

The pseudo code of our algorithm for our ranking methodology is shown below.

- Given: T, a set of training articles represented as pairs of articles and positive indicator scores where a score of 0 indicates that the article contains no relevant data and a score of 1 indicates that the article contains relevant data. For example:

(article 1, 0)

(article 2, 1)

(article 3, 1)

(article 4, 1)

(article 5, 0)

(article 6, 1)

Articles 2, 3, 4, and 6 all are highly relevant to the content being scored for, and articles 1 and 5 do not have relevant content.

- Given: A method for determining that two words are equivalent, or equivalently, a method for normalizing the text in an abstract (i.e., removing non alphanumeric characters and making case uniform) so that the overall amount of unique words is reduced.

  For example: In the sentence "Peptide motif-binding functions for binding of SH3/SH2 domain containing proteins." Would normalize to "PEPTIDE MOTIF BINDING FUNCTIONS FOR BINDING OF SH3 SH2 DOMAIN CONTAINING PROTEINS".

The algorithm pseudo-code is as follows.

Generation of Word Scores from article summaries / training values.

- Define t, u, and v as maps where the keys are strings and the values are integers. The sum of the scores will be stored in u, and the number of times each word has appeared will be stored in v.
- For each article "a" and score "s" in T:

  For each word "w" in a:

Increment u[w] by s

Increment v[w] by 1

- Calculate the average score for each word:

  For each word "w" in u:

  t[w] = u[w] / v[w]

- Define x: a map of articles to scores
- Define y: a map where the keys are strings and the values are integers. This will be used to count the number of appearances of each word.
- For each article "a" in the test set to be scored:

  For each word "w" in "a":

  Increment y[w] by 1

- Using y and t, calculate the Pearson correlation coefficient for a
- Set x[a] equal to (Pearson correlation coefficient for a / number of words in a)

Contained in x are the scores for the papers in the test set. Higher scores indicate a greater likelihood of relevance with respect to the content positively scored in the training set.