2015

Combining Disparate Data Types: Protein Sequences and Protein Structures

Kejue Jia
Iowa State University, kjia@iastate.edu

Robert L. Jernigan
Iowa State University, jernigan@iastate.edu

Combining Disparate Data Types: Protein Sequences and Protein Structures

Abstract

With the development of high-throughput, next-generation sequencing and other advanced technologies, a large number of **gene expression** profiles have been produced. Many of these profiles are available from public databases [1-3]. A challenging research problem that has drawn a lot of attention in the past is to infer gene regulatory networks from the expression data. A **gene regulatory** network is represented by a directed graph, in which nodes represent transcription factors or mRNA with edges showing transcriptional regulatory relationships between two nodes.

Disciplines

Biochemistry, Biophysics, and Structural Biology | Bioinformatics | Genetics and Genomics | Molecular Biology

Comments

This article is published as Kejue Jia and Robert L. Jernigan (2015) Combining Disparate Data Types: Protein Sequences and Protein Structures. J Data Mining Genomics Proteomics 6:e117. doi: 10.4172/2153-0602.1000e117. Posted with permission.

Data Mining in Genomics & Proteomics

Kejue and Robert, J Data Mining Genomics Proteomics 2015, 6:1
http://dx.doi.org/10.4172/2153-0602.1000e117

Editorial Open Access

Combining Disparate Data Types: Protein Sequences and Protein Structures

Kejue Jia and Robert L. Jernigan*

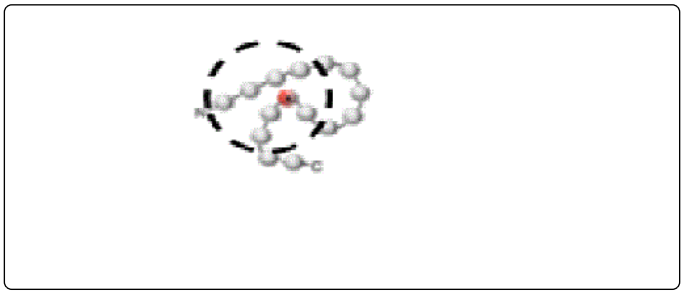
Bioinformatics and Computational Biology Program, Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University, Ames, IA 50011

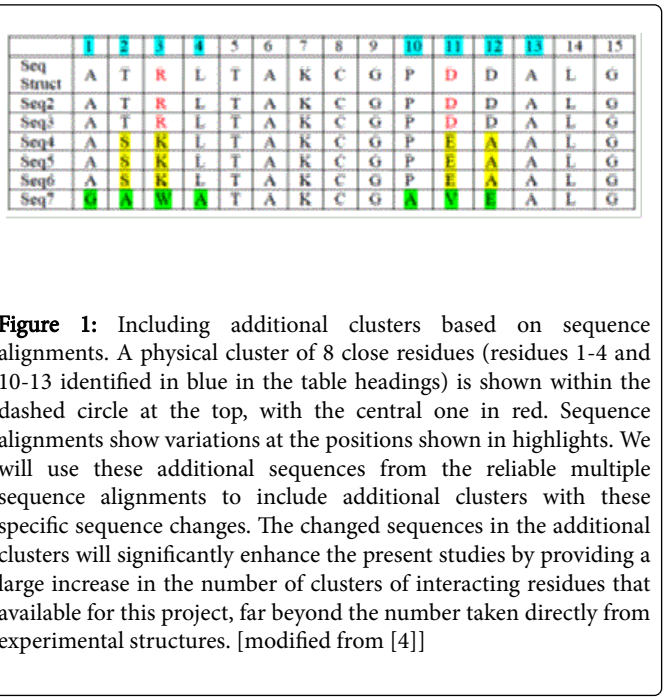
***Corresponding author:** Robert L. Jernigan, Bioinformatics and Computational Biology Program, Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University, Ames, IA 50011, Tel: 1-515-294-3833; E-mail: jernigan@iastate.edu

Rec date: Dec 25, 2014; **Acc date:** Dec 26, 2014; **Pub date:** Jan 2, 2015

Editorial

There are many opportunities from combining disparate types of data. The example below is the investigation of protein variability and evolution by combining protein structures with sequences. The many large-scale genome sequencing projects and the advent of individual organism and metagenome sequencing is starting to accumulate in the enormous numbers of protein sequences. In some cases there are tens of thousands of sequences related to a single protein. Together with the 100,000+ structures in the Protein Data Base (PDB), this remarkable data for comprehending the important structural and sequence relationships. Understanding sequence conservation is obviously important for the understanding of protein evolution and ultimately for understanding phenomics. The datamining opportunities are unprecedented for using these available big data sets to develop a deeper understanding of protein evolution. We pioneered such approaches with protein structures in 1985 by extracting potentials for interacting amino acids when we were able to use only 42 protein structures, which were sufficient for extracting the counts of the 190 types of amino acid pairs [1]. These large new data await clever new applications by dataminers. One of our new projects uses these data to identify closely interacting tight clusters of amino acids to characterize their sequence and geometric variabilities. Amino acid substitutions in proteins can be significantly better understood by considering the closely interacting groups of amino acids within structures, which have been combined naturally for favorable collective multibody interactions tight packing. Two amino acids that are distant in sequence may fold up into close contact pairs in the native structure. Because they are close, if one of them is replaced with a smaller amino acid, one of its neighbors may be replaced by a larger one, to maintain protein stability. In densely packed proteins, these correlated relationships involve more than simply pairs. In our project, information derived from Multiple Sequence Alignments (MSA) will be used to expand the numbers of physical clusters taken from structures by substituting the amino acids, according to the sequence alignments (see Figure 1).





	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Seq Struct	A	T	R	L	T	A	K	C	G	P	D	D	A	L	G
Seq2	A	T	R	L	T	A	K	C	G	P	D	D	A	L	G
Seq3	A	T	R	L	T	A	K	C	G	P	D	D	A	L	G
Seq4	A	S	K	L	T	A	K	C	G	P	E	A	A	L	G
Seq5	A	S	K	L	T	A	K	C	G	P	E	A	A	L	G
Seq6	A	S	K	L	T	A	K	C	G	P	E	A	A	L	G
Seq7	G	A	W	A	T	A	K	C	G	A	V	E	A	L	G

Figure 1: Including additional clusters based on sequence alignments. A physical cluster of 8 close residues (residues 1-4 and 10-13 identified in blue in the table headings) is shown within the dashed circle at the top, with the central one in red. Sequence alignments show variations at the positions shown in highlights. We will use these additional sequences from the reliable multiple sequence alignments to include additional clusters with these specific sequence changes. The changed sequences in the additional clusters will significantly enhance the present studies by providing a large increase in the number of clusters of interacting residues that available for this project, far beyond the number taken directly from experimental structures. [modified from [4]]

By applying the sequence alignment to generate a larger number of possible clusters, we will be directly including evolutionary information. In a protein, amino acids co-evolve with other amino acids in ways to compensate for changes that are introduced. Characterizing these from a large set of proteins will permit understanding these interactions better. In multiple-sequence alignments of a given protein from different biological sources these co-evolving residues can be identified. Even the intricacies of allostery and how the proteins move and respond to other molecules could be meaningfully investigated with these large sets of data. These groups of correlated mutations can give insights into the structure and function of a protein.

Contact clusters from a set of PDB structures can be selected to have different CATH topologies. The CATH database [2] is a classification of protein structures from Protein Data Bank. It contains a semi-automatic, hierarchical classification of protein domains. The four main levels in classification are Class, Architecture, Topology and Homologous superfamily. Protein structures that have same topology level share particular structural features. The current version of CATH database (version 4.0) includes 69,058 annotated PDBs. There are alternative ways to obtain MSA, from Pfam, or by using different multiple sequence alignment procedures, such as, MUSCLE and CLUSTAL Omega. Pfam [3] is a database of curated protein families, each of which is defined by two alignments and a profile hidden

Markov model (HMM). Protein sequences within one family are aligned according to their functional regions, commonly termed domains. The current version of the database is Pfam 27.0, which contains a total of 14831 families. Table 1 shows a partial list of Pfam families and the number of sequences in each family.

Family ID	# Sequences	Family ID	# Sequences
PF00023	10811	PF00376	2639
PF00042	2007	PF00439	10075
PF00067	78448	PF00515	10038
PF00071	7863	PF00620	9481
PF00104	9214	PF01047	12725
PF00233	3100	PF01381	30271
PF00348	15289	PF01966	22279

Table 1: Pfam protein family ID and the number of sequences in each multiple sequence alignment

The structural clusters and their sequences will capture the complex evolutionary information from the sequence alignments. The phylogenetic information could even be used as a weighting scheme for the clusters. The tight clusters can be quite specific, and such clusters will no longer depend just on the types of pairs of amino acids involved, but rather on larger pieces of structure, i.e., they will be protein-specific (different clusters for different proteins). Thus, a protein-specific cluster set can be derived separately for each protein. Previously Sander, Marks, and Onuchic have succeeded in predicting structural contact pairs of amino acids from the sequence data [4-6]. By utilizing the strength of the inferred couplings, they developed predictors of residue-residue proximity that have proven useful for protein structure prediction. The multi-body clusters described above provide a significantly more cooperative representation than do pairwise clusters, and also show impressive gains in threading calculations. Thus, we expect them to be superior for distinguishing the importance of specific clusters.

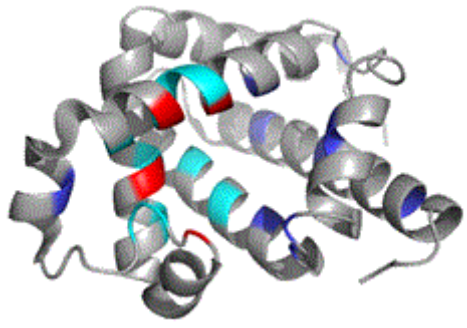


Figure 2: Myoglobin structure showing three sets of functionally related amino acids, marked in red, cyan and blue color, identified by their residue substitution patterns.

When the amino acids are no longer required to be spatially close to one another. These conserved groups of amino acids from a MSA are collectively correlated with protein functions. Myoglobin is shown in Figure 2 (PDB:2mgm). Three sets of amino acids are highlighted in red, cyan and blue. They are connected to the function of the protein. The amino acid substitutions in MSA may suggest novel mechanisms for collective behaviors of such disparate sets of amino acids to achieve certain functions, which may not have been known from previous research, and understanding these become more important as more proteins are utilized as drugs.

Overall the interpretation of the sequence data becomes significantly more meaningful when they are combined with structural information. There are many important opportunities in datamining by combining diverse data types.

References

1. S Miyazawa, RL Jernigan (1985) Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 18:534-552.
2. Sillitoe I, Cuff AL, Dessailly BH, Dawson NL, Furnham N, et al. (2013) New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. Nucleic Acids Res 41: D490-498.
3. RD Finn, A Bateman, J Clements, P Coggill, RY Eberhardt, et al. (2014)The Pfam protein families database. Nucleic Acids Res; 42 (Database issue):D222-D230.
4. DS Marks, LJ Colwell, R Sheridan, TA Hopf, A Pagnani, et al. (2011) Protein 3D structure computed from evolutionary sequence variation. PLoS One 6:e28766.
5. DS Marks, TA Hopf, C Sander (2012) Protein structure prediction from sequence variation. Nat Biotechnol 30:1072-1080.
6. Morcos F1, Jana B, Hwa T, Onuchic JN (2013) Coevolutionary signals across protein lineages help capture multiple protein conformations. Proc Natl Acad Sci U S A 110: 20533-20538.