RESEARCH PAPER Asian Journal of Bio Science, Vol. 4 Issue 2 : 274-277 (October, 2009 to March, 2010)

Functional annotation of histone proteins in human

MANOJ KUMAR SHARMA[1], ASTHA DUHAN[2] **AND** M.L. SANGWAN[3]
[1]Department of Bio-informactics, J.V. College, Baraut, BAGHPAT (U.P.) INDIA
[2]Department of Bio-informatics, Mata GujriCollege, FATEHGARHSAHEB (PUNJAB) INDIA
[3]Department of Animal Biotechnology, C.C.S. Hissar Agricultural University, HISSAR (HARYANA) INDIA

(Accepted : September, 2009)

Proteins are very much useful components and their functional annotation is also very useful as well. The function of protein is the measure of the expression of that particular protein. By knowing the function of one protein it can be found out the function of that protein also which has conserved region of the above protein whose function is known. PANDORA is a web based tool to aid biologist in interpretation of protein sets without the need of examining each individual protein. The general approach that PANDORA uses is based on annotation. In PANDORA, annotations are treated as binary properties that can be assigned to proteins. In relation with histone protein family we find out not only functional annotation but the evolutionary relationship with the family members of histone protein. PANDORA gives results for histone protein family. It provides more white the nodes which mean the sensitivity is higher, that are close to 1, reflect the result that fraction of the proteins with annotation. Specificity provides the data that is always more than 0, that gives the result that fraction of protein set has annotation.

Key words : PANDORA, Histone

INTRODUCTION

Proteins are complex nitrogenous organic biopolymers of amino acids showing great diversity in their organization and they are of prime biological importance. Proteins are the most complex chemicals synthesized in nature and must fold into complicated three-dimensional structures to become active. Family and super family classification also serves as the basis for rule-based procedures that provide rich automatic functional annotation among homologous sequences and perform integrity checks. Patterns or profiles, numerous rules have been defined to predict position-specific sequence features such as active sites, binding sites, modification sites, and sequence motifs. Linking protein data to more bibliographic data that describes or characterizes the proteins is crucial for increasing the amount of experimental information and improving the quality of protein annotation. The annotation of function by transference from proteins of related sequences is not the only possibility for the "in silico" prediction of function. The flourishing of genomic data has enabled other modes of function prediction independent of the identification of homologous sequences. The function of proteins can be inferred from the study of the similarity of their expression pattern with properties of a system can be explained by but not deduced from its components (such as protein domains).

Biological reality actually indicates just the opposite; the presumption that fold similarities alone are sufficient to identify functional similarity is discredited in numerous cases (Koppensteiner Devos D, Koppensteiner, Skolnick, Karplus , Tramontano, Fischer, Kolinski ,Rost, Flockner , Jones, Kelley, Rychlewski, Skolnick, Valencia A., 2000). Methods of annotation on the basis of sequence similarity (such as BLAST (Smith TF, Zhang X.,1997)) or sequence motifs (such as Blocks, PRINTS, Pfam, and Prosite) have proven successful, they are limited by implicit assumptions underlying their methodology. A number of new sequence analysis challenges have emerged in the genome era. Predicting the function of each newly found protein has been a main focus of genome analysis (Bork P, Dandekar T, Diaz-Lazcoz Y, Eisenhaber F, Huynen M, Yuan Y., 1998). A general approach for functional characterization of unknown proteins is to infer protein functions based on sequence similarity to annotated proteins in sequence databases. This complex and ambiguous process is inevitably error prone (Bork and Koonin, 1998).

Histones are highly conserved proteins that serve as the structural scaffold for the organization of nuclear DNA into chromatin. The histones have an amino terminal tail, a globular domain, and a carboxy-terminal tail. Histone H1, the most common form of linker histone, binds to nucleosomal DNA at the point from which the DNA exits the nucleosome, and is required for higher order packing of chromatin. The four core histones, H2A, H2B, H3 and H4 assemble into the octamer (2

molecules of each). Histones are modified post-translationally by the actions of enzymes in both the nucleus and cytoplasm that deposit specific functional groups. These modifications help to regulate the processes that depend on DNA, such as transcription, DNA repair, recombination and replication.

PANDORA (Protein ANnotation Diagram ORiented Analysis):

It is a web tool based on the SwissProt protein database that allows us to carry out integrative biological annotation analysis of protein sets, using annotations from various sources. PANDORA currently integrates annotations from the following sources: SwissProt keywords, NCBI Taxonomy , InterPro, GO, SCOP and ENZYME.

MATERIALS AND METHODS

First retrieved histone sequences from NCBI (National Center for Biotechnology Information), *i.e*. http://www.ncbi.nlm.nih.gov

The URL for PANDORA is:-http://www.pandora.cs.huji.ac.il/

PANDORA for functional annotation of protein:

There are mainly five members in the histone family, *i.e*. H1, H2A, H2B, H3, H4 . As an exception there can be a sixth type also, that is H5. They all share some conserved part in them. In the experiment we have taken two members together and found out the result of those. There we get 15 output result files. Then we go for the same for three members together, that gives 20 output result files, for four and five members, they give 15 and 6 (Table 1, 2, 3, 4, 5 and 6) output results, respectively. They also give graph, that give details of protein at a particular node the color of the node gives the sensitivity of that particular node protein.

The input to PANDORA is a protein set and a selection of one or more annotation types. The system displays the full protein-keyword relations between the proteins of the set and the keywords of the selected types. This is displayed as an intersection-inclusion Directed Acyclic Graph (DAG). An intersection-inclusion DAG is a hierarchical graph that describes all intersection and inclusion relationships between given sets. In our case, these sets would be protein sets, each protein set sharing a unique mixture of keywords. This allows presentation of the whole collection of protein-keyword relations without loss of the initial information.

Construction of the graph:

The annotations on your protein set as a binary matrix where the rows are annotations and the columns are your proteins. Each row describes a subset of proteins that share a certain biological property. Each of the subsets is the basic nodes of the graph. PANDORA compares these nodes and constructs a hierarchical graph of them. Each node represents a set of proteins that were all assigned a common annotation. When comparing two nodes there are three possible cases:

Sets are equal:

Nodes will be merged into one set of proteins. The annotation that will be assigned to this node will be that of both parents.

One set is a subset of the other set:

An edge will be created between the two nodes, and the subset will be placed beneath.

Sets intersect (excluding the previous case):

A new node will be created containing the intersection of the two sets. The annotation that will be assigned to this node will be that of both parents. Edges will be created from the two nodes two the new node and it will be placed beneath them.

Sets are disjoint:

Leave as two separate nodes.

The graph:

Nodes in the graph (appear as red and white balls) represent sets of proteins, sharing a unique combination of annotations. Their size is relative to the amount of proteins in them. To see the annotations given to the proteins of a node, move the mouse pointer over a node. The edges (appear as green lines) represent subset/superset relations between the nodes, with a top-to-bottom directionality. This means that if node “A” is connected to node “B” which is beneath it, “A” is a superset of “B”. This provides a simple yet important rule to follow: each of the proteins of a node shares its annotations and the annotations of all its ancestors in the graph. The node at the top of the graph represents all the proteins of set, even if the proteins do not share any annotation (in this case it will be marked as “BS” - Basic Set). Clicking this node will open a window that lists the protein of this set. Clicking any other node will show the proteins of that node as a new graph. To view the proteins of any other node in the graph, first click on it, and in the new window that opens click on the top node. The more white the

node is, the higher its sensitivity (a completely white node has a sensitivity close of around 1, and a completely red node has a sensitivity of around 0). (Fig. 1)

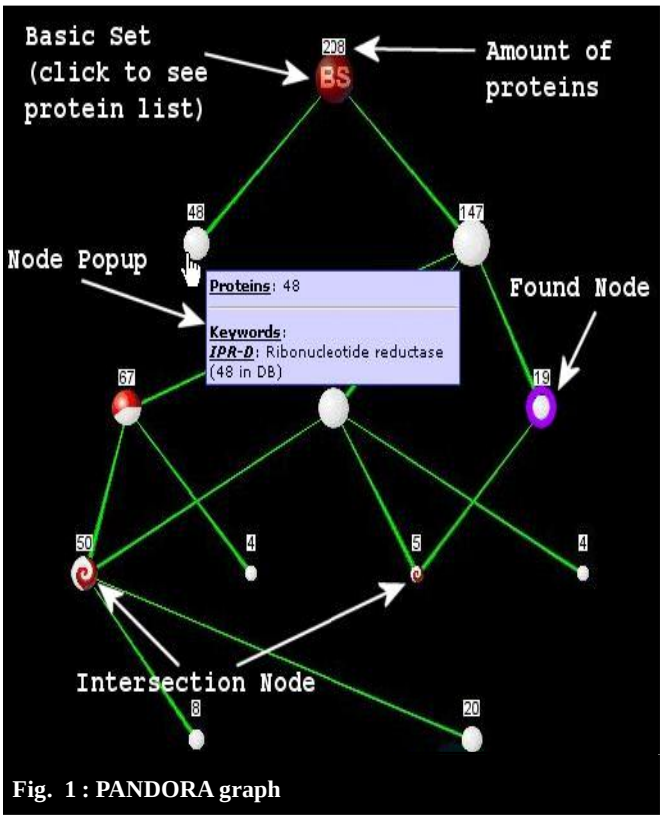

Fig. 1 : PANDORA graph

RESULTS AND DISCUSSION

The results obtained from the present investigation are summarized in Table 1, 2, 3, 4 and 5.

PANDORA results:

For Five members together

All of the tables shows the inter relation between the members of the family in terms of sensitivity and specificity. It shows that sensitivity is higher, that are close to 1, reflect the result that fraction of the proteins with

Table 1 : H1-H2A-H2B-H3-H4

Sr. No.	Keyword	Amount	Sensitivity	Specificity
1.	Histone H3	663	1	0.408
2.	Histone H1/H5	317	1	0.195
3.	Histone H2A	270	1	0.166
4.	Histone H5	242	0.883	0.149
5.	Histone H2B	201	1	0.124
6.	Histone H4	173	1	0.107

Table 2 : H1-H2A-H2B-H3-H5

Sr. No.	Keyword	Amount	Sensitivity	Specificity
1.	Histone H3	663	1	0.384
2.	Histone H1/H5	559	1.763	0.324
3.	Histone H5	516	1.883	0.299
4.	Histone H2A	270	1	0.157
5.	Histone H2B	201	1	0.117

Table 3 : H1-H2A-H2B-H-4-H5

Sr. No.	Keyword	Amount	Sensitivity	Specificity
1.	Histone H1/H5	559	1.763	0.453
2.	Histone H5	516	1.883	0.418
3.	Histone H2A	270	1	0.219
4.	Histone H2B	201	1	0.163
5.	Histone H4	173	1	0.14

Table 4 : H1-H2A-H3-H-4-H5

Sr. No.	Keyword	Amount	Sensitivity	Specificity
1.	Histone H3	663	1	0.391
2.	Histone H1/H5	559	1.763	0.329
3.	Histone H5	516	1.883	0.304
4.	Histone H2A	270	1	0.159
5.	Histone H4	173	1	0.102

Table 5 : H1-H2B-H3-H-4-H5

Sr. No.	Keyword	Amount	Sensitivity	Specificity
1.	Histone H3	663	1	0.407
2.	Histone H1/H5	559	1.763	0.343
3.	Histone H5	516	1.883	0.317
4.	Histone H2B	201	1	0.123
5.	Histone H4	173	1	0.106

Table 6 : H2A-H2B-H3-H-4-H5

Sr. No.	Keyword	Amount	Sensitivity	Specificity
1.	Histone H3	663	1	0.419
2.	Histone H5	274	1	0.173
3.	Histone H2A	270	1	0.171
4.	Histone H1/H5	242	0.763	0.153
5.	Histone H2B	201	1	0.127
6.	Histone H4	173	1	0.109

annotation and shows the members of the family are very much related in terms of conserved regions present. Specificity provides the data that is always more than 0, that gives the result that fraction of protein set has annotation. It shows that histone protein family is annotated through the members of the family. If any family has specificity 0, the members of the family do not share similarities during the course of evolution.

REFERENCES

Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M., Yuan, Y. (1998). Predicting function: from genes to genomes and back. *J. Mol. Biol*., **283** : 707–725.

Bork, P. and Koonin, E.V. (1998).Predicting functions from protein sequences:where are the bottlenecks?. Nature Genet 1998;18:313–318. Acad Sci USA;100:581–586.

Devos, D., Koppensteiner, Skolnick, Karplus, Tramontano, Fischer, Kolinski, Rost, Flockner, Jones, Kelley, Rychlewski, Skolnick and Valencia, A. (2000). Practical limits of function prediction. *Proteins*, **41**:98–107.

http://www.ncbi.nlm.nih.gov

http://www.pandora.cs.huji.ac.il

Smith, T.F. ande Zhang X. (1997).The challenges of genome sequence annotation or "The devil is in the details". *Nature Biotechnol*., **15** : 1222–1223.