VOL.6 NO.2 September 2017 - ISSN 2233 – 1859

Operations Research Society in Bosnia and Herzegovina

Southeast Europe Journal of Soft Computing

Available online: http://scjournal.ius.edu.ba

IUS Soft Computing Research Group

Inferring Protein Function from Structure

M. Ljubijankic

Faculty of Engineering and Natural Sciences,
Department of Biological Sciences and Bioengineering
International University of Sarajevo International University of Sarajevo,
HrasnickaCesta 15, Ilidža 71210 Sarajevo,
Bosnia and Herzegovina
maidaaa-k@hotmail.com

Article Info

Article history:

Article received on 17 June 2017

Received in revised form 17 August .2017

Keywords:

Structure; function; prediction.

Abstract

A major goal of molecular biology is to understand functions of all genes in nature. Accordingly, it is of great importance to improve large-scale functional genomics and proteomics experiments. However, due to costly and time-consuming nature of experiments, bioinformatics approach to infer the function appears to be very attractive. Besides this, there are many proteins of known structure which are not yet functionally characterized. This makes the investigation of sequence-function and structure-function relationships even more necessary. The number of methods for in silico annotation of function has increased enormously over the past few decades, from methods that rely on high sequence similarity between a protein of unknown function and a family of well-characterized proteins to methods that rely on "profiles" to infer the function. Although computational approach of inferring protein function is an important challenge, there are many obstacles to overcome. First, a function is not well defined and can be defined at several levels of detail. Accordingly, it is very difficult to create controlled vocabularies. Second, the precise values for thresholds of significant sequence similarity are actually specific to particular aspects of function and have to be re-established for any given task. The most common approach to study the function is through evolutionary relationship, or homology, with proteins of known function and it is based on the assumptions that "homologous proteins that have similar sequences and structures, have similar functions" which is the so called *Sequence-Structure-Function Paradigm*. In this research project, the limitations of this approach are studied.

1. INTRODUCTION

Although the possession of sequence similarity is indicative of underlying structural similarity and may impose similar function, it is not always the case. Sometimes it can be the case that homologous proteins have different function and different amino acid sequence (Sangar et al., 2007). This threatens to produce incomplete or even erroneous annotations if the annotation is passed freely among homologues.

As described by Whisstock and Lesk (2003), several mechanisms of protein evolution that produce altered or novel functions include (1) Divergence, (2) Recruitment and (3) Mixing and matching of domains.

Divergence occurs when a few specific mutations modulate specificity of closely related proteins in a family. For example, the trypsin family of serine proteinases contains a specificity pocket. Mutations tend to leave the backbone conformation of the pocket unchanged but affect the shape and charge of its lining, altering the specificity. Additionally, variation in the residues responsible for mediating catalysis may lead some enzymes to include novel molecules as their substrates. In some cases, very large divergence has led to very different function. Whisstock and Lesk (2003) described the relationships between homology, structure and sequence divergence, and functional change. They identified following changes in folding pattern, or topology, which are associated with functional changes

1. Addition/deletion/substitution of secondary structural elements
2. Circular permutation
3. Stand invasion and withdrawal
4. Changing the topology while maintaining the architecture

Recruitment generally represents an action of protein adopting a new function with no sequence change. An interesting example is the duck eye lens proteins or phosphoglucose isomerase which directly recruited to a new role by modification of gene expression without prior gene duplication (Whisstock and Lesk, 2003; Wistow, 1993).

Mixing and matching of domains may represent an obstacle in functional annotation because a vast majority of proteins are multi-domain proteins. Since domains may possess independent functions, modulate one another's function, or act together to provide a single function, the function prediction of multi-domain proteins may be a great challenge (Whisstock and Lesk, 2003).
Another important aspect to consider in function annotation is a degeneracy of the genetic code. Since more than one triplet codon encodes the same amino acid, a gene may acquire silent mutation (not changing the encoded amino acid) or expressed mutation (changing the encoded amino acid). Change in the amino acid may lead to different structure thus contributing to the different function (Benner et al., 2000). For example, mutations in amino acid positions may change the substrate specificity of enzymes, what further may lead these enzymes to be replaced by ones which are specific for that particular substrate (Schulz and Schirmer, 1979).

2. INFERRING PROTEIN FUNCTION FROM SEQUENCE-STRUCTURE SIMILARITY

Since Anfinsen's experiment until today - the exact sequence-structure-function relationship is not completely understood. As mentioned in the above, the most common approach to study and predict the protein function is through homology, that is, by detection of similarity of amino-acid sequences and/or structures by database search, and assuming that the proteins identified as homologues are with similar functions (Whisstock and Lesk, 2003). Anfinsen's experiment is miss interpreted by the biology common sense and concluded that the amino acid sequence of a protein is able to determine its structure. In turn, it is concluded that the biochemical function of a protein is determined largely by its 3D structure (Tramonto, 2005). The existence of many counter-examples that do not follow the "'sequence -> structure -> function" 'rule, caused the emergence of a new research area.
Many examples led us to conclude that sequentially homolog proteins have the similar structures. However, based on this assumption, we cannot conclude that nonhomolog proteins always have different structures, because, very often, apparently sequentially unrelated proteins share a similar topology. Also, there are proteins that share similar structures with no recognizable relationship between the sequences and vice-versa, proteins with similar sequences but with the dissimilar structures (Tramonto, 2005; Whisstock and Lesk, 2003).
Accordingly, questioning the*sequence-structure-function* paradigm may be a big challenge in computational function annotation.

2.1 Questioning The*Sequence-Structure-Function* Paradigm

What makes it possible to assume that similar sequences have similar functions is the fact that protein function is often carried out by a set of specific conserved amino acids which often come in the form of a pattern. For example, residues forming active sites or binding specific ligands. These kind of patterns are known as 'deterministic patterns'. On the other hand, a 'stochastic pattern' reports the probability that one amino acid occupies a certain position (Tramonto, 2005). Regardless of the pattern used, the main question remains: What are sequence similarity measures/thresholds for the safely transferring function between related proteins? (Lee, Redfern and Orengo, 2007). Moreover, nature provides us with examples of recruitment (described above), but also with examples where underlying sequence similarity doesn't imply functional similarity. Several studies (Devosand Valencia, 2000.; Wilson et al., 2000.; Todd et al., 2001.; Rost, 2002; Tian and Skolnick, 2003; Sangar et al., 2007; Addou et al., 2008) in the past few decades have investigated this issue and tried to elucidate the sequence-function relationship.
Since structure changes much more slowly than sequence through mutations, structural information can provide more accurate function prediction than sequence-based methods. This approach relies on the existence of superfamilies of proteins, described as a set of homologous proteins with similar 3D structures and related, though not necessarily identical, biochemical functions (Petsko and

Ringe, 2004). Moreover, functional sites generally occupy well-conserved topological locations in the structure and sometimes, even with no detectable homology between proteins that share the same fold, these functional sites still tend to locate in the similar regions of the 3D structure (Rigden, 2009).

As Rentzsch and Orengo (2013) explained, there exist many more pairwise homology relationships between individual domain sequences than between whole-protein sequences.

The basic approach to studying protein function in this manner is by identification of certain domains or folds. However, again, there is an important question to answer: ''How similar must two proteins be at the structural level to have similar functions?'' Lee, Redfern and Orengo (2007) stated: ''An analysis of the CATH database revealed that although most domains that share the same fold are associated with a single function, a small number of 'superfolds' (such as the ubiquitous Rossmann fold) can be associated with upwards of 50 different functions. Furthermore, these superfolds are the most common folds and account for >50% of domain sequences with predicted structures.'' The situation is even more dramatic in superfamilies that exhibit significant structural divergence; there may be an insertion of secondary structure elements that tend to colocate to produce surface features that change the active site or promote different protein-protein interactions.

Besides this, a great challenge represents the identification of domain boundaries in amino acid sequence, what leaves methods and algorithms for domain identification with no reliable set of positive examples on which they can be trained. Moreover, domains in proteins are not necessarily contiguous in sequence; they may start folding in a particular domain, then form another domain (Tramonto, 2005).

Additionally, there are pieces of evidence of homologous proteins that adopt different folds or multiple, changeable folding motifs depending on time and conditions. This further may have consequences on function prediction (Rigden, 2009). Besides this, the fact that structure wise very closely related protein families can have completely different biological functions may impede the function prediction. A well-known example of crystallins describes how even at high levels of sequence identity there are significant differences in function. In this case, crystallins retain more than 50% sequence identity to enzymes, but function as structural proteins in the eye lens (Petsko and Ringe, 2004). Another example includes ABC transporter family in which structure for some members appear to be homologous, but the function is divergent (Luckie et al., 2003). These counter-examples and instances which escape from the rule ''structure->function'' represent the main problem in the structure-based function prediction.

Accordingly, our study aims to provide a clearer picture of the biggest challenges in computational function prediction. Moreover, it aims to shed a light on the sequence-structure-function relationship by reviewing the studies in order to understand basic concepts and ideas used to annotate the function correctly. Besides this, our goal is to analyze all currently available proteins in PDB which are functionally similar in order to establish homology and functional similarity between them.

Specific Aims

1. To analyze the most common challenges and problems in protein function prediction.

Since protein function represents the most important information not only to biologists but many other scientists as well, proper analysis and prediction are of the great importance for the scientific community. Therefore, when experimental methods for function annotation are not accessible and feasible, computational approaches are used to predict protein function. Due to obstacles that appear to hamper proper function prediction, it is of great importance to analyze and study those challenges and gain an insight into the mechanisms to overcome them.

2. To study sequence-structure-function relationship and its relation to protein function prediction**.**

The more technology improvements are enabling better computational methods, the more sequence-structure-function paradigm diverge from its meaning. More and more examples are found to escape from 'sequence implies structure implies function' rule and identification and analysis of those allow us to better understand conditions under which this rule won't hold and what are the basic concepts when using this relationship to predict protein function.

3. To investigate homology relationship between proteins in PDB and relation between homology and functional similarity of those proteins.

In order to gain a complete insight into homology and functional similarity relationship, in silico analysis of functionally similar proteins may be a reasonable step. Since several studies have found that not all homologous proteins share the similar structure and vice-versa, detailed in silico analysis of sequence alignments, domain and motif identifications may provide an additional information about sequence/structure homology and functional similarity relationship.

Background Results

Protein function prediction and annotation have become an emerging need in the past decade since novel protein structures are deposited in PDB at enormously big rates. Currently, there are 125310 protein structures in PDB database. Among them 3448 proteins (for which the structure is determined by X-ray) are of unknown function. These numbers suggest that function annotation, whether

by experimental methods or computational methods, is of great importance.

When it comes to computational methods, as discussed above, several challenges appear to impede the annotation process. The underlying cause for this is a complex relationship between sequence, structure, and function.

Sangar et al. (2007) studied quantitative sequence-function relationships in proteins based on gene ontology and obtained interesting results. More specifically, they studied the relationship between divergence of sequence and function in homologous proteins, using the Molecular Function DAG of Gene Ontology for the classification of function. In their study, they explained that when it comes to Enzyme Commission (EC) classificationand the relationship between sequence similarity and functional similarity, several authorsreachedsimilaroptimistic conclusions. Most authors agree that at levels of sequence identity > 40%, precise function is conserved and some reported that approximately 90% of pairs of proteins with sequence identity > 40% conserve all four EC numbers (for the summary of several studies, see Table 1.) Devos and Valencia (2000) also reported the ability to predict correctly the agreement of Families of Structurally Similar Proteins (FSSP) categories and SWISS-PROT keywords, as a function of the level of sequence similarity.

Table 1. Summary of studies exploring sequence identity thresholds

Dataset size (number of protein sequences/ families/subfamilies)	Dataset composition	Conclusions	References
2338	Homologous pairsof PDB structuraldomainsorsequences	50% identity is required for conservation of all four EC digits	Devos and Valencia, 2000.
n/a	29,454 representativepairs of structural domains from SCOP.Comparisonsatvarious levels of the SCOP hierarchy(family, superfamily,fold)	40% identity is required for conservation of all four EC digits	Wilson et al., 2000.
65303	Homologous pairsof structural domains from CATH and theirsequence relatives from SwissProtand GenBank.	40% and 60% identity is required for conservation of all four EC digits in single- and multidomain proteins	Todd et al., 2001.
26243	Whole proteinsequences with anEC annotation.	Below 70% identity, both the first and the fourth EC digits start to diverge	Rost, 2002.
22645	Whole proteinsequence homologues including enzymesand non-enzymes(derived byPSI-BLAST).	40% (60%) identity is required for conservation of the first three (all four) EC digits	Tian and Skolnick, 2003.
7868 protein families	Protein families with different numbers of members and withseed and full alignments of proteins in each family.	Between 40% and 60% sequence identity shows the highest change in the identical functions. The threshold at about 40% sequence identity, at which the observed behavior changes	Sangar et al., 2007
721 enzyme superfamilies and 3210 enzyme functional subfamilies	3210 enzyme functional subfamilies identified in 721 CATH-Gene3D enzyme superfamilies.	For more than 60% sequence identity, proteins share the same EC number in 90% of cases	Addou et al., 2008.

Sangar et al. (2007) investigated divergence of functions within the same "branch" of the DAG (those for which the lowest common ancestor of two nodes was not the root node) and those in different "branches" of the DAG. These functions are called similar and dissimilar functions, respectively.

As an example, they used EF-hand family and found out that as the sequences progressively diverge, there is a systematic decrease in the number of pairs with distance 0 (identical function). For 80–100% sequence identity the distribution of similar functions has a unique peak at 0.

Their data suggest the interesting result that there is a threshold at about 40% sequence identity, at which the observed behavior changes. As they stated:‘’For pairs of proteins with 0–40% residue identity, the distribution is largely independent of sequence identity. Above 40% sequence identity, there is a significant increase in similar functions over dissimilar ones’’ (Sangar et al., 2007).

When they analyzed combined PFAM data (the PFAM families were divided into five categories according to size: 2–30 members (5834 families), 31–60 members (719 families), 61–270 members (244 families), 271–780 members (27), and > 780 members (3 families)), the results show that between 40% and 60% sequence identity shows the highest change in the identical functions and concluding that functional divergence by mechanisms other than recruitment generally requires > 40% amino acid substitution. Moreover, they observed an example of

recruitment (the peak at distance = 6, for the proteins with 81–100% sequence similarity).

Rost et al. (2003) in their paper analyzed automatic prediction of protein function and explained that accuracy in transferring function requires several steps to be fulfilled. As stated in the paper, those are:

1. Build data sets that have experimental annotations about the presence (true, e.g., all proteins experimentally known to be nuclear) and absence (false, e.g., all proteins experimentally known not to be nuclear) of a certain aspect of function.
2. To avoid estimates that are incorrectly biased by the distribution of today's experimental information, extract a representative subset of proteins from the true data and align it against all proteins in the true set (minus the representative subset) and false set.
3. For all alignments, count how many true and false we find at every given threshold for sequence similarity.

The authors also investigated the level of accuracy in annotation process and concluded that >60% pairwise sequence identity is required for a transfer withless than 30% errorsand for errors below 10%, >75% sequence identity (Rost et al., 2003).

These thresholds are important to consider when annotating the function because incorrect functional assignment can easily propagate to new database sequence entries and undermine the value of protein annotation (Addou et al., 2008). As Pearson (2013) emphasized, inferring functional similarity based solely on significant local similarity is less reliable than inferences based on global similarity and conserved active site residues.

When it comes to structural ‘'thresholds" in transferring the function, there are several studies which focused on domain analysis to infer the function. Addou et al. (2008) were exploring the extent to which similarity between domains can be used to accurately infer functional information (Table 1). Their results on whole protein sequences comparison level were similar to previously published studies. However, on the domain level, they observed slightly increased numbers: 70% (50%) sequence identity is required for 90% confidence in matching full (third EC level) annotations.

When it comes to domains and multi-domain architecture (MDA) and their relation to function conservation among enzyme homologues, data suggest that if two relatives share similar domain partners, the likelihood that these two relatives (the proteins) have the same function is increased (Table 2). As Addou et al. (2008) observed, information on domain architectures can lower levels of pairwise sequence identity at which safe functional annotation can be transferred. This is important for functional annotation of uncharacterized sequence relatives as these relatives often lie within the twilight zone of sequence identity (<35%).

Based on the findings for EC annotation transfer, authors were interested to determine what transfer levels could be achieved when including non-enzyme sequences. Results show that 80% sequence identity is required for 90% transfer confidence, as opposed to 50% for EC transfers.

Whisstock and Lesk (2003) reviewed in details function prediction and difficulties and successes related to it. They concluded that many folds are compatible with very different activities. The five most ‘versatile’ folds are the TIM barrel, axb hydrolase, the NAD-binding fold, the P loop-containing NTP hydrolase fold, and the ferredoxin fold. However, there are several folds that appear in combination with only one function, which appears to be significant for function prediction.

Table 2.Generic sequence identity thresholds to be met for transferring functions with 100% and 90% confidence (adopted from Addou et al., 2008).

Function conservation level (i.e., confidence level for transferring functions)	Minimum sequence identity (identical MDAs) (%)		Minimum sequence identity (MDAs unknown) (%)	
	Sequence	Domain	Sequence	Domain
100% conservation of all four EC digits	80	90	80	90
>90% conservation of all four EC digits	60	60	60	70
100% conservation of the first three EC digits	70	80	70	80
>90% conservation of the first three EC digits	40	40	40	50
Note: Thresholds differ for third- and fourth-level EC conservation and can be lower for domain-based comparisons when the proteins' MDA is known (underlined).				

Experimental Design

Aim 1: To analyze the most common challenges and problems in protein function prediction.

- A short literature review of the most common challenges related to protein function prediction such as pairwise sequence alignment, multiple sequence alignment and function vocabulary. In this review, we will explain general problems and try to describe the optimal solutions which could be introduced in solving them.

Aim 2: To study sequence-structure-function relationship and its relation to protein function prediction.

- A literature review of sequence-structure-function relationship and the relation between homology and functional similarity. This aims to provide detailed information about sequence identity values and protein structure information which are required to safely transfer the function between proteins. Here, the goal is to summarize and list the results from recent studies on homology and functional similarity in order to gain an insight into the rules which have to be followed to annotate the function properly.

Aim 3: To investigate evolutionary relationship between proteins in PDB and relation between homology and functional similarity of those proteins.

- Based on the commonly used homology analysis algorithms, we aim to predict the level of homology between functionally similar proteins in PDB. For this purpose, several different analysis and tools will be utilized, since we aim to correlate sequence and functional similarity, but also the structural and functional similarity between the proteins. In order to obtain significant and reliable results, the same analysis can be done with more than one tool.
- Retrieval of protein sequences will be from pdb website (https://www.rcsb.org/pdb/home/home.do)

- To analyze sequence identity of functionally similar proteins by pairwise sequence alignment.

For this purpose, we will use both global and local alignment.

EMBOSS Stretcher (https://www.ebi.ac.uk/Tools/psa/emboss_stretcher/) calculates an optimal global alignment and uses a modification of the classic dynamic programming algorithm which uses linear space. The alignment maximises regions of similarity and minimises gaps using the scoring matrices and gap parameters provided to the program (McWilliam et al., 2013).

EMBOSS Water (https://www.ebi.ac.uk/Tools/psa/emboss_water/) uses the Smith-Waterman algorithm (modified for speed enhancements) to calculate the local alignment. A local alignment searches for regions of local similarity between two sequences and are very useful for scanning databases when you wish to find matches between small regions of sequences, for example between protein domains. Water finds an alignment with the maximum possible score where the score of an alignment is equal to the sum of the matches taken from the scoring matrix (Bleasby, 2009).

- To analyze sequence identity of functionally similar proteins by multiple sequence alignment.

We will align amino acid sequences by the standard dynamic programming algorithm using the BLOSUM 62 matrix. For this purpose alignment program, MUSCLE will be used.

MUSCLE (MUltiple Sequence Comparison by Log-Expectation) (https://www.ebi.ac.uk/ Tools/msa/muscle/) is an online tool for multiple alignments of protein sequences. Following guide tree construction, the fundamental step of this tool is pairwise profile alignment, which is used first for progressive alignment and then for refinement (Edgar, 2004).

- To analyze homology between functionally similar proteins.

In order to gain more knowledge about the homology-function relationship, we will try to establish homology between our set of functionally similar proteins. For this purpose, we aim to use HMMER (https://www.ebi.ac.uk/Tools/hmmer/) – a web server that uses profile hidden Markov models (HMMs) to represent the query which can take the form of a single protein sequence or a multiple sequence alignment. In our study, we will use multiple sequence alignment, for which the observed amino acid frequencies in each column are converted to position-specific probabilities, with per position probabilities for both insertions and deletions, determined from the input alignment (Finn et al., 2015).

BLAST (Basic Local Alignment Search Tool) (https://blast.ncbi.nlm.nih.gov/Blast.cgi) is a sequence similarity tool that uses heuristics to produce results and calculates an “expect value”, which estimates how many matches would have occurred at a given score by chance, what further can aid a user in judging how much confidence to have in an alignment (Madden, 2013).

- To analyze domains of functionally similar proteins.

Although different definitions and meanings of domain appear in the literature, general idea says that a domain constitutes a region of conserved sequence between different proteins, which may equate to a functional unit of a protein. Regardless of the definition, due to its ‘conserved’ nature, domain identification may aid the assessment of homology predictions but also provide an insight into the function (Emes, 2008).

In this study, we aim to use some commonly used domain identification tools to identify and compare domains which may be important for the functional similarity in our set of proteins. Conserved Domain Database (CDD) (https://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi) is a resource for the annotation of protein sequences with the location of conserved domain footprints, and functional sites inferred from these footprints (Bauer et al., 2015). It offers a way to find conserved domains for many protein queries, what is important for our study.

SMART (Simple Modular Architecture Research Tool) (http://smart.embl-heidelberg.de/smart/batch.pl) is another web server that allows the identification and annotation of genetically mobile domains and the analysis of domain

architectures. Domain detection in SMART relies on multiple sequence alignments of representative family members (Schultz et al., 2000).
PFAM (http://pfam.xfam.org/) is an online tool which allows comparison of single or multiple query sequences to libraries of HMMs. It is adatabase of curated protein families, each of which is defined by two alignments and a profile hidden Markov model (HMM) (Finn et al., 2014).

Expected Results

Analysis and review of homology-function and sequence-structure-function relationships should provide a better insight into challenges and obstacles present during protein function prediction process. Additionally, we expect that analysis of functionally similar proteins by different bioinformatics tools should help in better understanding of those relationships and generally, provide a clearer picture of those relationships.

Conclusions

The most important challenge in computational biology and bioinformatics is to understand and annotate the function to all proteins. Different approaches to function prediction are widely used, but the homology-based prediction is the most common one. Since it relies on the fact that similar sequence implies similar structure implies similar function, many algorithms are developed to analyze and elucidate those relationships. However, many obstacles impede the annotation process since there are many examples that represent the exception of this fact. Review of those obstacles would provide a better understanding of the problems and possible solutions, while in silico analysis of the homology-function relationship of functionally similar proteins would provide additional information and elucidation of this complex relationship.

REFERENCES

Addou, S., Rentzsch, R., Lee, D., &Orengo, C. (2009). Domain-Based and Family-Specific Sequence Identity Thresholds Increase the Levels of Reliable Protein Function Transfer. *Journal Of Molecular Biology*, *387*(2), 416-430. http://dx.doi.org/10.1016/j.jmb.2008.12.045

Benner, S., Chamberlin, S., Liberles, D., Govindarajan, S., & Knecht, L. (2000). Functional inferences from reconstructed evolutionary biology involving rectified databases – an evolutionarily grounded approach to functional genomics. *Research In Microbiology*, *151*(2), 97-106. http://dx.doi.org/10.1016/s0923-2508(00)00123-6

Bleasby, A. (2009). *Help - EMBOSS-Align*. *EMBL-EBI*. Retrieved 1 November 2017, from http://www.biomol.it/unictbiolmol-lab/figure_didattica/Help%20with%20Align.pdf

Devos, D., & Valencia, A. (2000). Practical limits of function prediction. *Proteins: Structure, Function, And Genetics*, *41*(1), 98-107. http://dx.doi.org/10.1002/1097-0134(20001001)41:1<98::aid-prot120>3.3.co;2-j

Edgar, R. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792-1797. http://dx.doi.org/10.1093/nar/gkh340

Emes, R. D. (2008). Bioinformatics: Structure, Function and Applications. *Methods In Molecular Biology™*, *453*. http://dx.doi.org/10.1007/978-1-60327-429-6

Finn, R., Bateman, A., Clements, J., Coggill, P., Eberhardt, R., & Eddy, S. et al. (2013). Pfam: the protein families database. *Nucleic Acids Research*, *42*(D1), D222-D230. http://dx.doi.org/10.1093/nar/gkt1223

Finn, R., Clements, J., Arndt, W., Miller, B., Wheeler, T., & Schreiber, F. et al. (2015). HMMER web server: 2015 update. *Nucleic Acids Research*, *43*(W1), W30-W38. http://dx.doi.org/10.1093/nar/gkv397

Lee, D., Redfern, O., &Orengo, C. (2007). Predicting protein function from sequence and structure. *Nature Reviews Molecular Cell Biology*, *8*(12), 995-1005. http://dx.doi.org/10.1038/nrm2281

Luckie, D., Wilterding, J., Krha, M., & Krouse, M. (2003). CFTR and MDR: ABC Transporters with Homologous Structure but Divergent Function. *Current Genomics*, *4*(3), 225-235. http://dx.doi.org/10.2174/1389202033490394

Madden T. (2013). The BLAST Sequence Analysis Tool. In: The NCBI Handbook [Internet]. 2nd edition. Bethesda (MD): National Center for Biotechnology Information (US). https://www.ncbi.nlm.nih.gov/books/NBK153387/

Marchler-Bauer, A., Derbyshire, M., Gonzales, N., Lu, S., Chitsaz, F., & Geer, L. et al. (2014). CDD: NCBI's conserved domain database. *Nucleic Acids Research*, *43*(D1), D222-D226. http://dx.doi.org/10.1093/nar/gku1221

Pearson, W. (2013). An Introduction to Sequence Similarity ("Homology") Searching. *Current Protocols In Bioinformatics*. http://dx.doi.org/10.1002/0471250953.bi0301s42

Petsko, G., &Ringe, D. (2009). *Protein structure and function*. Oxford [England]: Oxford University Press.

Rigden, D. (2009). *From protein structure to function with bioinformatics*. [Dordrecht]: Springer.

Rost, B. (2002). Enzyme Function Less Conserved than Anticipated. *Journal Of Molecular Biology*, *318*(2), 595-608. http://dx.doi.org/10.1016/s0022-2836(02)00016-5

McWilliam, H., Li, W., Uludag, M., Squizzato, S., Park, Y., &Buso, N. et al. (2013). Analysis Tool Web Servicesfrom the EMBL-EBI. *Nucleic Acids Research*, *41*(W1), W597-W600. http://dx.doi.org/10.1093/nar/gkt376

Pal, D., & Eisenberg, D. (2005). Inference of Protein Function from Protein Structure. *Structure*, *13*(1), 121-130. http://dx.doi.org/10.1016/j.str.2004.10.015

Sangar, V., Blankenberg, D., Altman, N., &Lesk, A. (2007). Quantitative sequence-function relationships in proteins based on gene ontology. *BMC Bioinformatics*, *8*(1), 294. http://dx.doi.org/10.1186/1471-2105-8-294

Schultz, J. (2000). SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Research*, *28*(1), 231-234. http://dx.doi.org/10.1093/nar/28.1.231

Schulz, G., & Schirmer, R. (1979). *Principles of protein structure*. New York: Springer-Verlag.

Tian, W., & Skolnick, J. (2003). How Well is Enzyme Function Conserved as a Function of Pairwise Sequence Identity?. *Journal Of Molecular Biology*, *333*(4), 863-882. http://dx.doi.org/10.1016/j.jmb.2003.08.057

Todd, A., Orengo, C., & Thornton, J. (2001). Evolution of function in protein superfamilies, from a structural perspective 1 1Edited by A. R. Fersht. *Journal Of Molecular Biology*, *307*(4), 1113-1143. http://dx.doi.org/10.1006/jmbi.2001.4513

Tramontano, A. (2005). *The ten most wanted solutions in protein bioinformatics*. Boca Raton: Chapman & Hall/CRC.

Whisstock, J., &Lesk, A. (2003). Prediction of protein function from protein sequence and structure. *Quarterly Reviews Of Biophysics*, *36*(3), 307-340. http://dx.doi.org/10.1017/s0033583503003901

Wilson, C., Kreychman, J., & Gerstein, M. (2000). Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *Journal Of Molecular Biology*, *297*(1), 233-249. http://dx.doi.org/10.1006/jmbi.2000.3550

Wrzeszczynski, K., Ofran, Y., Rost, B., Nair, R., & Liu, J. (2003). Automatic prediction of protein function. *Cellular And Molecular Life Sciences (CMLS)*, *60*(12), 2637-2650. http://dx.doi.org/10.1007/s00018-003-3114-8