A practical guide for the computational selection of residues to be experimentally characterized in protein families

Alfonso Benítez-Páez, Sonia Cárdenas-Brito and Andrés J. Gutiérrez

Submitted: 23rd May 2011; Received (in revised form): 2nd August 2011

Abstract

In recent years, numerous biocomputational tools have been designed to extract functional and evolutionary information from multiple sequence alignments (MSAs) of proteins and genes. Most biologists working actively on the characterization of proteins from a single or family perspective use the MSA analysis to retrieve valuable information about amino acid conservation and the functional role of residues in query protein(s). In MSAs, adjustment of alignment parameters is a key point to improve the quality of MSA output. However, this issue is frequently underestimated and/or misunderstood by scientists and there is no in-depth knowledge available in this field. This brief review focuses on biocomputational approaches complementary to MSA to help distinguish functional residues in protein families. These additional analyses involve issues ranging from phylogenetic to statistical, which address the detection of amino acids pivotal for protein function at any level. In recent years, a large number of tools has been designed for this very purpose. Using some of these relevant, useful tools, we have designed a practical pipeline to perform in silico studies with a view to improving the characterization of family proteins and their functional residues. This review-guide aims to present biologists a set of specially designed tools to study proteins. These tools are user-friendly as they use web servers or easy-to-handle applications. Such criteria are essential for this review as most of the biologists (experimentalists) working in this field are unfamiliar with these biocomputational analysis approaches.

Keywords: functional divergence; residue selection; multiple sequence alignment; mutagenesis; protein families; evolutionary information

INTRODUCTION

Combining computational and experimental approaches are the current trend in scientific research to solve central cell biology questions. This synergic effort permits a more rational design of scientific studies in favour of bench routines. The best biocomputational contributions to experimentalists are based on the development of tools for the selection and prioritization of the genes, proteins and residues to be examined. Consequently, a selection of candidates (i.e. genes, proteins, residues, etc.) based on rigorous analyses supported by biological and statistical evidence helps save the time and funds invested by wet laboratories. Protein characterization...
frequently involves dissection of residue(s) accomplishing a specific molecular task. In recent years, this field has been supported by computational studies addressing the identification of pivotal residues for a given protein function. Expert scientists in the study and characterization of a specific family of proteins normally obtain in-depth knowledge (after years of experimental research) of the particular and constrained set of residues, which helps classify a protein into a functional family (i.e. catalytic triads, dinucleotide binding motif, G-motifs of GTPases, DEAD-boxes, etc.). Thus, it is quite normal to find studies that examine the function of homologue residues in the different and/or novel members of the given family being examined (i.e. NTPases, oxidoreductases, methyltransferases, helicases, proteases, etc.). Consequently, many studies contain redundant data about equivalent and conserved residues, which carry out the same task in homologue proteins. Moreover, there are several studies that consist in testing the residues that play new molecular roles, such as protein–protein interaction, molecular switches, substrate specificity, etc. In all these studies, the main approach used by experimentalists for residue selection is multiple sequence alignment (MSA). However, the limited knowledge that scientists have about MSA parameter adjustment (i.e. substitution matrix, word, window, algorithm, etc.) can produce poor quality alignments, which do not prove very useful for residue selection purposes. Besides, the use of a low number of sequences in MSAs is also a frequent error, especially when a large family is being studied. Accordingly, residue conservation in a specific position of MSA will probably be true for that small set of analysed proteins, but not for a larger set. In order to overcome such failures and to present further useful computational tools for the selection of undisclosed residues for functional characterizations in any family of proteins, some important considerations in MSA analyses are explained, while other alternative MSA-based methods for the selection of functional residues in a protein are described below.

**MSA AS CORNERSTONE**

Most scientific knowledge on the biological understanding of living organisms is based on the comparison of sequences, structures, pathways, reactions, metabolites, etc., to infer functionality by homology. MSA is the simplest tool to detect residue conservation in a protein family, at sequence level, and to reflect the most important residues, as well as the probably indispensable amino acids for the correct function of that family of proteins. However, MSA can change according to the number of sequences, the number of organisms (and variability), the amino acid substitution matrix, and the algorithm being used. Thus, it is important to take into account certain considerations to use an appropriate MSA method.

**CHOOSING THE RIGHT MSA ALGORITHM**

Among the different algorithms designed for MSA, accuracy, scalability and computational cost must also be considered. Nowadays, the best method or algorithm to build an MSA must be capable of: (i) processing alignments based on dozens or hundreds of protein sequences; and (ii) being flexible in comparing non-homologue proteins sharing a functional domain. Since ClustalW was published in 1994 [1], it has rapidly become the most widely used MSA algorithm. However, some improvements have been made since then, and more accurate and faster methods have been developed. Some of the current, improved methods include MAFFT [2], PROBCONS [3], T-COFFEE [4] and MUSCLE [5, 6]. In particular, we should use highly accurate MSA methods (i.e. PROBCONS or T-COFFEE) given their implications for subsequent analyses. However, if computational cost is a disadvantage, MUSCLE offers good performance and global balance in accuracy, scalability and computational cost terms [7]. Like most MSA tools, it can be executed locally by downloading a simple file (for the Linux or Windows systems), and by calling the MSA algorithm using a basic command line; alternatively, we can upload an input text file of sequences on the MUSCLE web server at EBI (http://www.ebi.ac.uk/Tools/msa/muscle/), where we can use other reliable tools such as MAFFT, T-COFFEE or KALIGN to perform MSA. Both options retrieve an MSA output file in the FASTA format, which is readable by any MSA viewer.

**SEQUENCE SELECTION FOR MSA**

The characterization of the functional residues of a protein is frequently based on a poor MSA analysis with 3–5 homologue sequences (maximum 10) to
detect residue constraints. To further exploit MSA and the data acquired from it, we should build an MSA from a representative sample of the entire set of organisms in which the query protein is present. Thus, if we are interested in studying a bacterial family of proteins, we should have at least one representative protein sequence of each bacterial group in accordance with conventional taxonomy (i.e. Alfaproteobacteria, Betaproteobacteria, Delta/Epsilon, Gammaproteobacteria, Firmicutes, Cyanobacteria, Chloroflexi, etc.). To compile a useful file of sequences for MSA, non-redundant searching in major protein databases must be performed: UNIPROT [8] or GENBANK [9]; both are accessible using the BLAST tool [10] from the NCBI site [http://blast.ncbi.nlm.nih.gov/Blast.cgi]. In addition, we have to select a large number of sequences, preferably more than twenty-five (whenever possible), with the highest possible variability. For this object, the BLAST interface offers an advanced option to limit searches by taxonomy, which includes or excludes groups, families, genera or species. In this way, we should select a representative sequence per genus, or even per family, when organisms are highly related. If there is a low number of sequences to be analysed in the MSA, we could add related sequences by detecting distant homologues following the PSI-BLAST approach [11], construct MSAs with the PROMALS algorithm (http://prodata.swmed.edu/promals/promals.php) [12, 13], or modify the BLAST parameters by employing more permissive BLOSUM matrices [14], such as BLOSUM-45 instead of BLOSUM-62; the latter is used by default in BLAST searches. Alternatively, we can find additional members of the query family by searching in the Pfam database [15], which contains the largest collection of protein families, domains and motifs clustered by sequence conservation and amino acid profiles based on Hidden Markov Model (HMM). If we are interested in studying the residues of a functional domain in the query protein, we can include non-homologue proteins, but those that contain equivalent functional domains in their architecture. Thus, we can apply further phylogenetic and statistical analyses to the MSA, such as those described in the following sections. A quantitative evaluation of sequence selection and variability in MSA can be obtained from the DIVERGE analysis (see below).

**DETECTING FUNCTIONAL CONSTRAINTS IN MSA**

By employing user-friendly MSA viewers such as JALVIEW [16, 17] or SEAVIEW [18, 19], we can explore the MSA text file to search the highly conserved positions along the alignment. These conserved positions can be highlighted by colouring according to different criteria, such as percentage of identity. The residues showing whole conservation in the expanded set of sequences used for MSA will provide more information than those from an MSA based on a low number of sequences. In addition to residue conservation, the biochemical nature of residues can also be constrained. Consequently, different positively- (K, R) or negatively (D, E) charged residues will be present in a specific position within the MSA. Conservation is not so easy to observe in a polar or hydrophobic position. If we wish to draw a residue map for each position easy for our eyes and brains to understand, an amino acid profile based on the HMM can be built from MSA [20], interpreted by the HMM-Logo server [21] on its web site or locally by downloading the HMMVE viewer [22]. Good views of amino acid preferential usage per position in MSA can help us understand and recognize most of the relevant residues and those selected by evolution forces.

Complementary studies must be carried out if one of the query proteins has a three-dimensional X-ray or NMR structure; otherwise, we could follow several strategies to obtain a protein structure model in which we can localize the selected residues [23]. Thus, each conserved residue/position in the MSA can be localized on the respective protein structure, and some predictions of their respective roles may be stated for subsequent experimental corroboration. These predictions can focus on the residues adjacent to the active site of the query proteins, or even on the protein surface where they are possibly involved in protein–protein interaction (PPI) or in nucleic acid binding.

**BEYOND THE MSA ANALYSIS**

Further analyses can address the detection of those residues predicted to be involved in a specific task. One of these methods is the *In Silico* Two Hybrid method (2H), which predicts PPIs based on the amino acid-correlated changes/mutations between interacting proteins [24]. Moreover, other methods have been developed to detect altered functional
constraints, which directly compare different clusters (or subfamilies) of a protein family (or superfamily). In this brief guide, we focus on the last methodology because it permits MSAs to be explored in depth and to extract the evolutionary information deriving from speciation/specialization after gene duplication. The following methods have been illustrated in Figure 1 to provide a better understanding.

**WHAT DOES THE FUNCTIONAL DIVERGENCE ANALYSIS CONFER?**

The residue constraints extracted from an MSA can be altered if we add other functional-related proteins to the original MSA. These new proteins incorporated into this MSA must essentially be the paralogues that emerge after gene duplication during evolution. They are probably specialized in other molecular functions, but retain both high sequence identity and similarity to the original query protein(s) because of their closely related function. At this point, we have two different clusters belonging to a common evolution-related family, or superfamily, of proteins; a direct comparison made among all the sequences present in both clusters will provide valuable ways to detect those residues of functional relevance which are difficult to disclose by direct MSA observation in viewers, but are easier to detect through phylogenetic and statistical analyses [25–27]. The DIVERGE tool [28] (web site: http://www.xungulab.com) offers a complete interface to manage phylogeny, MSA and protein structure data. Its aim is to study type-I (or the site-specific rate shift of amino acids) and type-II

---

**Figure 1:** General pipeline for the selection of residue candidates for experimentally functional characterization. The steps and paths in the pipeline are outlined. There are steps that are connected by grey arrows in which, in some cases, the required computational tools are cited. The blue arrows in the boxes show three major computational methods to select residues (Functional Constraints, Functional Divergence and PCA), and show the localization of the hypothetical candidate residues to be analysed in each case.
(or the radical shift of the amino acid property) functional divergence [25, 27]. This methodology has provided good results in the different protein families to which it has been applied [29–31]. Although DIVERGE does not define a specific function for the residues selected in the query protein, it finds the residues that most probably act as functional determinants, which are ranked according to statistical evaluation. We can modulate such statistical significance by increasing the cut-off value (posterior probability score ≥0.85) to rescue residues under a strong functional divergence, together with a function assignment, in accordance with our expertise and structural data. The results obtained with this approach will largely depend on the MSA analysis where the selection of sequences and its variability are critical. For the quantitative proposes of this step, DIVERGE retrieves the coefficient of functional divergence (θ) after each pair-wise cluster comparison. As a result, a θ value higher than zero suggests that a significantly altered functional constraint has occurred between a pair of clusters. Then if θ does not differ from zero, this means that the divergence in MSA and in each cluster of proteins does not suffice to detect functional constraints. Therefore, we should attempt to make comparisons between different clusters or acquire new sequences to be studied.

DIVERGE results are also influenced by phylogenetic tree construction methods. In that way, a reliable and accurate phylogenetic tree have to be constructed prior to this analysis. Given that tree construction is essential for any DIVERGE analysis, we should follow a PROTTTEST approach [32] to determine the best evolutionary model for the set of query proteins. This web server (http://darwin.uvigo.es/software/prottest.html) only requires an MSA file to be used to test more than 60 different evolutionary models in the protein family. The selection of a specific model explaining molecular variation in this set of proteins will be based on ‘goodness-and-fit’ measures such as Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC), which will represent the uncertainty of all the models tested and the relevance of different model parameters (invariable sites: +I; amino acid categories: +G; amino acid frequencies: +F) in the evolution of query proteins. After testing all the different evolutionary models, PROTTTEST retrieves a rank of models according to the AIC or BIC values. Consequently, the evolutionary model with the lowest AIC or BIC value will be selected to draw the evolutionary tree of proteins.

Despite having obtained a reliable tree from the PROTTTEST for DIVERGE analysis, the DIVERGE suite offers a recommended option to enhance protein clustering when users are not familiar with tree presentation or if users are using an input tree that contains unsolved relationships. In such cases, the re-rooting of an input tree is a convenient way to select gene clusters before proceeding to the functional divergence analysis. By using the Neighbor-Joining option by selecting the ‘NJ Tree-Making’ button, the software will employ the Neighbor-Joining algorithm to quickly generate a tree based on the desired distance measure (i.e. p-Distance, Poisson or Kimura). Among the different distance calculation options, we should preferably employ those using corrected distances for multiple substitutions per site, such as the Poisson or Kimura algorithms. Once the tree has been re-rooted by DIVERGE, we can cluster all the proteins in well-defined groups, preferably with more than four sequences per cluster/group (i.e. bacteria versus mammals; or γ-proteobacteria versus α-proteobacteria for the proteins restricted to bacteria). At least two clusters are required to perform this analysis. However, if multiple clusters are selected, a pair-wise comparison should always be performed. Then we can compute types I and II functional divergence to extract the residues that are more likely to be under functional divergence.

In addition to the cluster comparison, a more specific intra-cluster analysis to detect functional divergence can be achieved according to phylogenetic distribution (see Supplementary Figure S1; where functional aspects of the characterised MnmC family of proteins [see ref 47–52] are outlined). Thus, a residue showing gain-of-function in some species within a particular analysed cluster can be easily observed by residue fixation in these sequences, but not in others belonging to the same protein cluster. This additional approach requires the extensive usage of highly related sequences instead of the diversity required for the above-mentioned cluster comparison. Furthermore, a cut-off value (or posterior probability) to select residues under functional divergence must be relaxed because the estimated variability of an amino acid per site is lower at this level of comparison.
COMPLEMENTARY ANALYSES

Despite having focused on a well-known method to detect the cluster-dependent amino acid conservation based on phylogenetic and statistical approaches, other alternative methods have been widely studied and successfully tested to functionally predict important residues. These methods include multivariate and related approaches that help predict pivotal residues when the functional and phylogenetic relationship is ambiguous [33–37]. Some methods explore the sequence information content of MSA to create a vector transformation where residues are evaluated according to their two- or three-dimensional clustering [33, 35, 37], while others use robust information from three-dimensional structures to predict the functional interfaces and residues involved in ligand-binding sites, protein–protein or protein–nucleic acid interactions, or ligand specificity [34, 36]. Notwithstanding, the gap between the available number of sequences and the three-dimensional structures of proteins can limit the use of the latter method type.

For the purpose of performing an additional analysis that is complementary to the functional divergence method extensively described as the scope of this guide, we can use the TreeDet web server (http://treedetv2.bioinfo.cnio.es/treedet/index.html) to predict the functional sites in a family of proteins [38]. TreeDet integrates related methods using a principal component analysis (PCA) to study the ‘Sequence Space’ of an input MSA. It also predicts functional sites after their representation in a multidimensional plot and statistical evaluation. The TreeDet authors encourage us to submit MSA from MUSCLE, T-COFFEE and PROBCONS algorithms, all of which are briefly described in earlier sections of this guide.

EXPERIMENTAL TESTING OF SELECTED RESIDUES

After completing the various steps stated throughout this guide for residue selection as regards functional characterization in a family of proteins, it is important to filter the set of selected residues for experimental testing. Accordingly, we should reduce the number of possible protein mutants to be studied by the directed mutagenesis approach. Then, we should focus on those residues with greater statistical and biological significance, otherwise the design and management of several mutants can result in a tedious and expensive experimental task that is time-consuming, and can even lead to ambiguous results, which are difficult to explain by the multiple mutants and functionalities, explored.

It is also necessary to consider appropriate amino acid changes. Traditionally, alanine scanning has been the initial approach followed to study the relevance of a specific residue; nonetheless, the function of a mutant protein will notably fluctuate in accordance with the amino acid selected to replace the targeted one. Consequently, electrostatic charge changes will help better elucidate the role of a positively- or negatively charged residue [39]. Likewise, polarity changes shed light on the role played by hydrophobic residues [40]. PPIs, or pocket binding sites, could also be studied by applying the aforementioned criteria. Nevertheless, steric hindrance by replacing small side-chain amino acids with those that have bulky side-chains can impair PPIs with single amino acid substitutions [41, 42]. For this case, we may consider that the protein interactions involving other larger molecules such as DNA or RNA could be supported by several interaction sites. Therefore, multiple mutants will be necessary to notably impair interaction [43], even when considering steric hindrance in all the mutated positions.

EVALUATION OF SELECTED MUTATIONS

At this point, we refine a list of residues that are probably involved in functions such as catalysis, co-factor binding, proton donation and in the mediation of the PPIs obtained by applying MSA, functional divergence, phylogenetic and/or PCA-related methods. Moreover, another analysis can be done to describe the thermodynamic implications of the desired mutational changes on protein structure stability before proceeding to mutagenesis and experimental studies.

The CUPSAT server [44] (http://cupsat.tu-bs.de) is a tool that predicts changes in protein stability upon point mutations. The CUPSAT evaluation of protein mutants requires the input files from PDB, as well as custom-developed protein structures. Its prediction model uses: (i) protein–environment potentials to predict protein stability; (ii) atom potentials; using the 40 amino acid atom classes from the Melo-Feytmans model [45, 46]; (iii) the torsion angles \( \varphi \) and \( \psi \); and (iv) the Gaussian apodization
function to adjust torsion angle perturbation in mutants. As a result, we can obtain a friendship list of the most probable and favourable amino acid changes given in ‘overall stability’, ‘torsion’, and ‘protein thermodynamics’ terms. Consequently, at the end stage of this analysis, we will have obtained strong computational evidence for both residues and probable mutations to support an experimental and practical approach to study the functional relevance of the different residues in any family of proteins.

SUPPLEMENTARY DATA
Supplementary data are available online at http://bib.oxfordjournals.org/.

Key points
- The considerations, issues, tools and suggestions reviewed in this guide have been addressed to improve the functional characterization of residues in any family of proteins under study by experimentalists.
- The general and central considerations of the MSA analysis have been reviewed for the purpose of constructing better and more informative alignments of protein families; simultaneously, different tools have been exposed according to user requirements.
- The different approaches presented herein help to shed light on the extraction of the evolutionary and functional information derived from sequence comparisons. A methodical usage of these approaches will permit a more rational design of experimental studies, which will be supported by stronger biological and statistical evidence.
- The aim of this review is to also promote the development of reliable, useful and combined computational bench routines given that the design of studies with more biological and statistical evidence prior to experimental testing could save time and funds, thus enabling more efficient research.
- Finally, the aim of the step-by-step design of this workflow is to encourage experimental biologists to learn and to acquire in-depth knowledge on biocomputational analyses, tools and methods which could permit flexible analyses according to user demands and the complexity of the query proteins without potential black box results assumed by the users.

Acknowledgements
The authors thank the Centro de Investigación y Desarrollo en Biotecnología for its support in the implementation, edition and publication of this guide; they also thank M.E. Armengod, I. Moukadir and M.J. Garzón for inviting us to collaborate in their MmuC project. Finally, they particularly thank the editors and peer reviewers of this guide for their constructive criticisms that have helped improve the quality of this review.

FUNDING
This study was supported by Centro de Investigación y Desarrollo en Biotecnología – CIDBIO in Bogotá DC, Colombia.

References


50. Pearson D, Carell T. Assay of both activities of the bifunctional enzyme MnmC that catalyses the last two steps in the biosynthesis of hypermodified nucleoside mmm5s2U in tRNA. *FEBS J* 2009;276:4169–83.
