OXFORD

Briefings in Bioinformatics, 17(2), 2016, 255–261

doi: 10.1093/bib/bbv045
Advance Access Publication Date: 2 July 2015
Paper

Practical analysis of specificity-determining residues in protein families

Mónica Chagoyen, Juan A. García-Martín and Florencio Pazos

Corresponding author: Florencio Pazos, National Centre for Biotechnology (CNB-CSIC) c/ Darwin, 3. 28049 Madrid, Spain. Tel.: +34.915854669; Fax: +34.915854506; E-mail: pazos@cnb.csic.es

Abstract

Determining the residues that are important for the molecular activity of a protein is a topic of broad interest in biomedicine and biotechnology. This knowledge can help understanding the protein's molecular mechanism as well as to fine-tune its natural function eventually with biotechnological or therapeutic implications. Some of the protein residues are essential for the function common to all members of a family of proteins, while others explain the particular specificities of certain subfamilies (like binding on different substrates or cofactors and distinct binding affinities). Owing to the difficulty in experimentally determining them, a number of computational methods were developed to detect these functional residues, generally known as 'specificity-determining positions' (or SDPs), from a collection of homologous protein sequences. These methods are mature enough for being routinely used by molecular biologists in directing experiments aimed at getting insight into the functional specificity of a family of proteins and eventually modifying it. In this review, we summarize some of the recent discoveries achieved through SDP computational identification in a number of relevant protein families, as well as the main approaches and software tools available to perform this type of analysis.

Key words: protein function; protein functional specificity; protein design; specificity-determining position (SDP); protein functional site; multiple sequence alignment (MSA)

Introduction

Obtaining the amino-acid sequence of a protein is relatively easy, compared with the difficulty of obtaining its three-dimensional (3D) structure or other experimental functional information. This is leading to an exponential increase in the number of protein sequences stored in public databases, which is orders or magnitude higher than the number of proteins whose structures are known or for which we have functional clues. Recent improvements in de novo sequencing and re-sequencing technologies [1, 2] are boosting this trend.

One way of taking advantage of this massive amount of data is to use sequence comparison methods to collect and compare homologous proteins (those sharing a common ancestor). Such a comparative study of the members of a group of homologous proteins (also termed 'superfamily') provides a lot of information on the functional and structural features of its members [3]). It is well established that homologous proteins share the same global 3D structure and many functional aspects (inherited from the common ancestor). Consequently, sequence comparison is most commonly used to predict the 3D structure and function of proteins, simply transferring these features from the homologous proteins for which they are known.

Apart from these global structural and functional analyses, protein sequence comparison allows us to study the pattern of conservation of individual residues. The first step to perform such a comparative study at the residue level is to generate a

Mónica Chagoyen is a postdoctoral researcher and staff technician at the Spanish National Center for Biotechnology (CNB-CSIC). Her research interests include functional bioinformatics, integrated data analysis and biological data management. She also provides bioinformatics support for experimental groups, leading the Sequence Analysis and Structure Prediction facility of the CNB.
Juan A. García Martín is currently a PhD student at the Biology Department of the Boston College. He has participated in the development of software for handling protein functional residues and collaborated with experimental groups providing then with bioinformatics support.
Florencio Pazos is staff scientist at the Spanish National Center for Biotechnology (CNB-CSIC), where he leads the Computational System Biology Group. His research is focused on the analysis of biological networks, especially metabolic networks, the prediction of protein functional and binding sites and the prediction of protein–protein interactions.

Submitted: 1 April 2015; **Received (in revised form):** 12 June 2015. **Accepted:** 15 June 2015

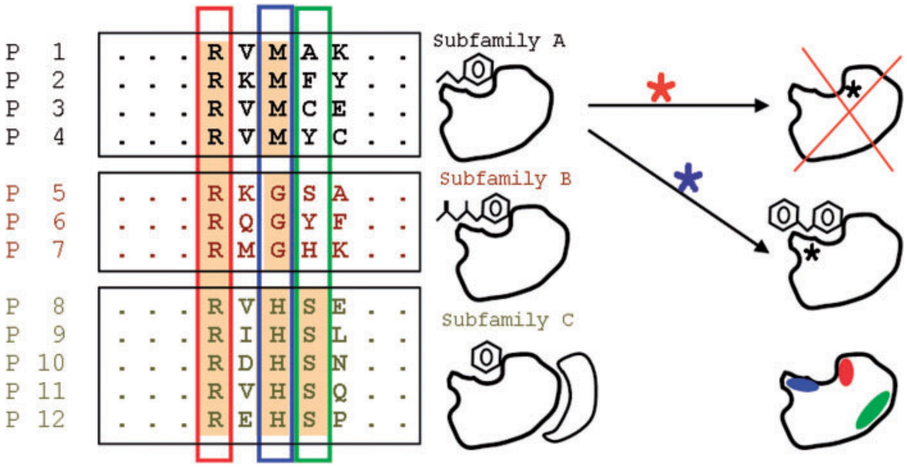


Figure 1. Representation of an idealized MSA with 12 homologous proteins, which can be grouped in three subfamilies with different functional specificities: they bind substrates of the same class but with some chemical differences (additionally subfamily C binds another protein). Besides the MSA, a representation of the (similar) structure of one member of each subfamily is shown. Three positions with different conservation patterns are highlighted in the alignment: full conservation (red, left highlighted column) and two types of subfamily-specific conservation (SDPs) (blue and green, middle and right highlighted columns). The mapping of these three types of positions in a generic structure of a member of the family is shown on the bottom right. Conservation owing to structural reasons is not shown here for clarity. For a member of family A, two mutations were designed (*): one involving a fully conserved position (red, top), which probably inactivates the protein, and another in an SDP position (blue, middle), which can lead to changed specificity. A colour version of this figure is available at BIB online: http://bib.oxfordjournals.org.

multiple alignment of the sequences of a group of homologous proteins ('multiple sequence alignment', MSA). In this alignment, evolutionary equivalent residues are stacked together (in the same column in the most common representation) (Figure 1). We can think of the MSA as a representation of the results of an 'experiment' in which evolution performed many trial/error cycles trying different amino-acid types at different positions, and keeping those 'mutations', resulting in functional proteins. Thus, a column in a MSA (termed 'position' in this context) can be regarded as a representation of the mutations 'allowed' by evolution in the corresponding residues of the homologous proteins. Consequently the mutational patterns of the positions in a MSA provide important functional information on the corresponding residues [4]. For example, a position with an invariable amino acid ('conserved position', Figure 1) is representing a protein site where no changes were allowed: an important site for the structure and/or function of the protein. Thus, conserved positions (Figure 1) were the first and most obvious predictors of functionally important residues from sequence information alone [5]. Being able to computationally predict the residues of a protein that are essential for its function is very important, taking into account how difficult and expensive it is to determine that experimentally. The knowledge of a protein's functional sites is important not only to understand its function at the atomic level, but also to devise ways to modify it.

Apart from full conservation, there are other patterns in MSA positions indicative of functional importance. If we can subdivide our protein family in different subgroups with distinct functional specificities (i.e. differing in certain functional details while sharing the global common function of all the proteins), some positions might show up with a subgroup-dependent conservation pattern: they are conserved only in particular subgroups, or the conserved amino acid is different for the different groups (Figure 1). The fact that their conservation is restricted to a particular subgroup suggests that their functional role is related to the functional specificities/peculiarities of that group, instead of the global function of the family. For example, within an alignment of enzymes catalyzing the same reaction but on different substrates, fully conserved positions would be pointing to the residues involved in the catalysis (the common function of the family) while group-dependent conserved positions could point to residues involved in the specific recognition of the different substrates (Figure 1). For these reasons, these positions are generally termed 'specificity-determining positions' (SDPs) and they complement the fully conserved positions described above as predictors of functionally important residues from sequence information alone. The functional information they provide is different from that of fully conserved positions. They can be used, for example, to design mutations to change the substrate specificity of an enzyme while keeping its catalytic activity, overall regulation and other functional aspects (Figure 1). Thus, they allow playing with subtler functional aspects of the proteins, whereas mutations in fully conserved positions would generally lead to a nonfunctional protein.

This review provides an overview of current computational approaches to detect SDPs in protein sequences and their potential applications in biotechnology and biomedicine. We will start with a number of published examples where the computational detection of SDPs was a fundamental part of a larger work (generally including experimental parts) aimed at better characterizing protein families of industrial/medical importance. They illustrate the main scenarios in which the study of SDPs can be useful. After that section, aimed at convincing the reader of the importance of SDPs, we will summarize the main computational methodologies for detecting them, focusing on those that are available to nonexpert users through graphical and web-based interfaces.

Examples of SDP analyses

Being able to discern which residues within a protein might be in charge of controlling its functional specificity can be useful in certain situations. In general, it allows getting insight into the atomic basis of that specificity, which is essential to design mutants with interchanged or new specificities, which has obvious applications in Biotechnology and Biomedicine.

Because the methods for detecting SDPs need many sequences of members of the family of interest to work, they were first applied to families that were the target of huge sequencing efforts (in the pre-genomic era) owing, for example, to their medical interest. The 'Ras' superfamily of small GTPases comprises a large number of eukaryotic proteins involved in several cellular functions [6, 7], including signaling cascades related to cancer. These proteins change their conformation in a cycle driven by GTP binding and hydrolysis and interact with different effectors. The Ras superfamily comprises >10 subfamilies, with particular functional specificities and playing different cellular roles. The evolutionary and functionally unrelated interactor(s) recognized by each subfamily determines its cellular role, this being as diverse as control of membrane traffic ('Rab' subfamily), signaling cascades involved in cell proliferation (Ras) or nuclear transport (Ran). The residues involved in GTP binding and hydrolysis are fully conserved in the superfamily, as this is the function common to all its members, while the positions responsible for the binding of the different effectors tend to show up as SDPs. Consequently, a number of studies tried to extract those SDPs from MSAs of this superfamily to better understand how these proteins work, as well as design mutants with altered/switched specificities. For example, Bauer *et al.* identified the two main SDPs of this superfamily and were able to swap the interaction specificities of the 'Ral' and 'Ras' subfamilies. Mutating these two residues was enough for making 'Ral' behave as 'Ras', while mutating only one of them they could make a 'Ras' protein function as a 'Ral' [8]. In this case, the 'SequenceSpace' method was used for detecting the SDPs (see next section).

Another family of biomedical interest comprises the 'carnitine/choline acyl transferases', involved in fatty-acid transport across membranes. Their functional specificities can be defined according to multiple axes: some of them use 'choline' as antiporter, while others use 'carnitine'; some can be inhibited by the presence of 'malonyl-CoA', while others are insensible to it, i.e. they transport fatty acids with different chain lengths, etc. By computationally extracting the residues associated to each of these specificities (SDPs), it was possible to design mutants with one of these functional aspects altered while the others remain unchanged. Examples include turning a malonyl-CoA-insensitive enzyme into another that is inhibited by this molecule, and the other way around [9], or switching the fatty-acid length preference, making an enzyme that normally transports short-chain fatty acids capable of transporting long-chain ones and the other way around [10]. In these cases, the authors used the 'SequenceSpace' program, complemented with interactive visual examination of the protein alignments and structures.

Also of biomedical interest is the family of receptors of psychoactive bioamines. These G-protein-coupled receptors respond, among others, to the endogenous hormones 'dopamine' and 'serotonine'. A computational analysis of the SDPs of this family allowed predicting the residues responsible for discriminating these two hormones [11]. The authors also discovered a second set of SDPs unexpectedly far from the hormone-binding site, and found that they control the efficacy in triggering the conformational change leading to G-protein activation. This site is allosterically connected with the hormone-binding site and opens interesting possibilities for drug development. Moreover, it was possible to convert a dopamine receptor into a serotonine receptor playing with mutations in these SDPs. The authors used the 'evolutionary trace' (ET) method (see next section) for performing the SDP analysis in this family of proteins.

In another interesting work, Ratnikov and co-workers detected the SDPs in a family of matrix metalloproteinases [12]. These enzymes degrade proteins at the extracellular matrix, being essential in processes such as morphogenesis, wound healing and tissue repair and remodeling. They are involved in the progression of diseases such as atheroma, arthritis, cancer and chronic tissue ulcers [13]. For the analysis of SDPs the functional subgroups were automatically defined (see next section) according to the patterns of cleavage efficiencies in a panel of phage-displayed peptide substrates. That is, proteins with similar cleavage patters were grouped together. As in other cases, it was possible to change the substrate preferences of these proteases mutating some of these SDPs, opening interesting possibilities for designing proteases with 'a-la-carte' peptide recognition patterns. Owing to the peculiarities of the subfamily classification (based on patterns of peptide cleavages), the authors used an ad hoc approach for performing the SDP analysis.

A similar analysis of SDPs was carried out with the 'JDet' package (see next section) in the family of bacterial amino acid racemases, a group of proteins of biotechnological interest [14]. The aim was to get insight into the residues controlling the range of substrate specificity of these enzymes: some of these enzymes are specific for alanine, while others are able to racemize a much larger spectrum of amino acids. Again, it was possible to experimentally demonstrate the involvement of these SDPs in substrate binding, opening the possibility of designing racemases with desired substrate specificities.

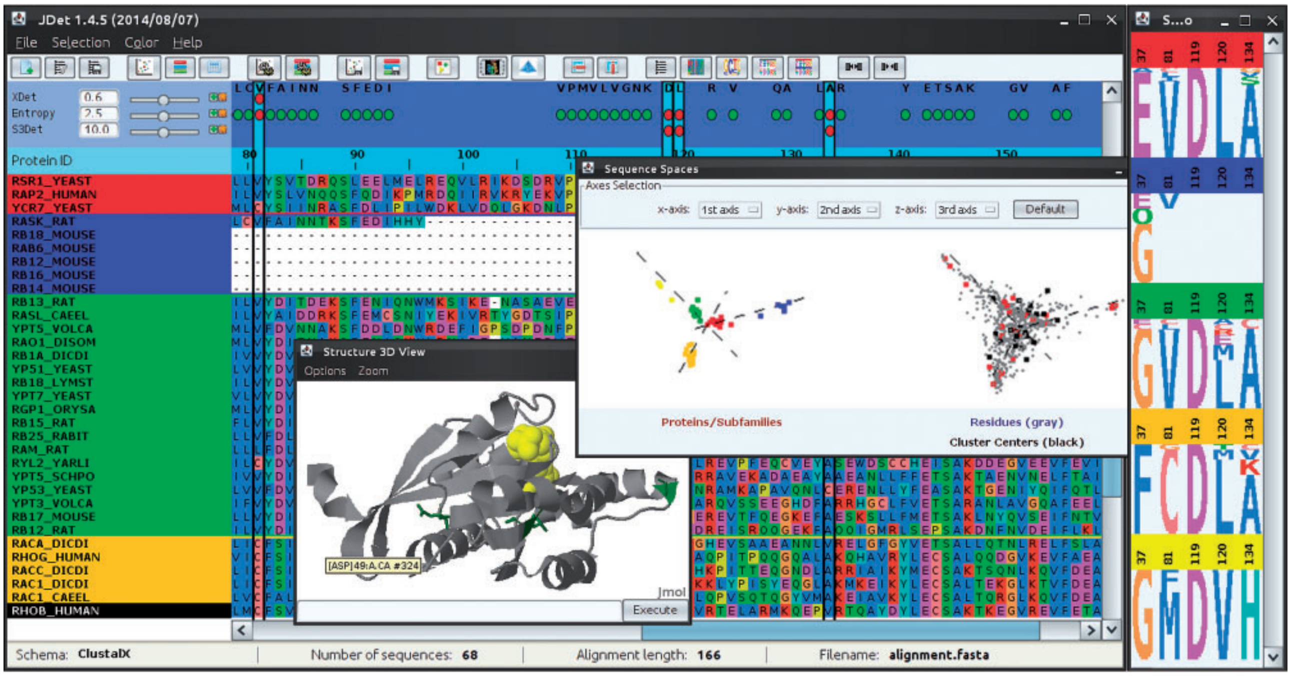
Another example is the sequence analysis of the eight subunits of the eukaryotic chaperonin 'CCT/TRiC'. This macromolecular complex assists in the folding of key proteins such as actin and tubulin. Its functional form is a hetero-octamer of eight different (albeit homologous) subunits forming a barrel inside which the refolding process takes place, in a cycle driven by ATP binding and hydrolysis. In prokaryotes, this molecular machine is composed of eight identical subunits, and they act in a concerted way during the ATP-driven folding cycle (the barrel is symmetric in its functioning). On the contrary, it is well known that in the eukaryotic CCT the subunits do not act in a concerted manner and they define an asymmetric barrel with distinctive functional sides. Nevertheless, the atomic determinants of this asymmetry were not known. An SDP analysis of the subunits, performed with the JDet package, allowed detecting the positions differentially conserved in them. Most of these positions were around the ATP binding site, suggesting that the different affinities for ATP of the subunits were the ultimate responsible for the functional asymmetry observed in this complex [15]. This observation also opens the possibility of designing CCT mutants with different degrees of asymmetry (or totally symmetrical) and studying their effects on its chaperone activity.

A family of bacterial 'small multidrug transporters' was also subject to SDP analysis, in this case using an ad hoc protocol, for locating the residues responsible for the binding of certain drugs (and hence responsible for the resistance to them). As in the other cases, by experimentally mutating these positions, the authors were able to change the specificity of the transporters what was reflected in the resistance profiles of the corresponding bacteria [16].

As a final example, in a recent study Chevalier *et al.* identified a mutant of *DWARF14*, an alpha/beta hydrolase involved in 'strigolactone' signaling during plant growth, revealing a residue essential for its function [17]. This position was in fact one of the SDPs that explained the specificity for 'strigolactone', as compared with other homolog enzymes acting on a different hormone signaling pathway.

Computational approaches for detecting SDPs

Although idealized representations such as that shown in Figure 1 can give the impression that the problem of locating SDPs in MSAs is easy, in the case of real alignments of large



Schema: ClustalX | Number of sequences: 68 | Alignment length: 166 | Filename: alignment.fasta

Figure 2. Screenshots of the interface of JDet. The main window contains the MSA loaded by the user, where five subfamilies (red, blue, green, orange and yellow) were automatically detected by the program. The SDPs detected by different programs are shown at the top row (green balls), and the corresponding positions in the MSA (columns) highlighted. The sequence logos for these positions are shown on the right. A window with a representative 3D structure with the SDPs highlighted on it and another with 3D projections of the sequence and residue spaces are shown. A colour version of this figure is available at BIB online: http://bib.oxfordjournals.org.

proteins with hundreds or even thousands of sequences it is far from being trivial. The main reason is that the conservation patterns in these real alignments are not perfect, and there are always deviations from the ideal trend illustrated in Figure 1. These deviations (non-perfect conservation) can be owing to different reasons, including sequencing errors and errors in the alignment. Also, allowing or not allowing conservative substitutions (changes involving chemically similar amino acids) can drastically change the conservation score of a given position. Indeed, even detecting fully conserved positions is far from being trivial [5]. In addition, there is the problem of distinguishing subfamily-specific conservation patterns that are related to function from those owing to other reasons (e.g. imposed by the uneven distribution of sequences in the MSA) [18]. Finally, the definition of the subgroups (e.g. the three subfamilies in Figure 1) can be problematic too, as for alignments with hundreds of sequences it is impossible to have experimental information on the functional class of every protein. For this reason, most approaches for detecting SDPs also 'predict' concomitantly the functional classes of the proteins present in the MSA, generally grouping them based on phylogenetic criteria.

All these difficulties led to the development of many different approaches for detecting SDPs in MSAs. One of the first approaches to detect SDPs was the ET method [19]. ET takes as input a MSA and the phylogenetic tree of the proteins within it. Based on this tree, the method explores successive hierarchical partitions of the MSA into more specific subfamilies, and look for the conserved positions showing up at each partition. ET ranks the positions of the MSA depending on the partition where they become conserved: fully conserved positions appear at partition 0 (the whole MSA), in partition 1 appear the positions that are differentially conserved in the two main subfamilies within the MSA, and so on. From its initial implementation, ET was improved in different ways (e.g. [20]). An advantage of ET-related approaches is that they detect not only SDPs but also functional residues in general (including fully conserved). A drawback, at least in the initial implementations, is that it does not consider conservative changes when evaluating subfamily conservation.

Another family of methods use a criterion based on mutual information to detect SDPs, given a partition of the MSA into subgroups. One possibility for defining that partition of the MSA is to look for the 'optimal' partition according with some criterion, such as that rendering the 'best' set of SDPs [21]. Another possibility is to use the ortholog/paralogs definition to partition the MSA, under the assumption that groups of orthologs are functionally homogenous [22]. It is also possible, using heuristic approaches, to try exploring all possible subfamily groupings (not only those coherent with a given phylogenetic tree) and report that which maximizes some criterion, together with its associated SDPs. This is the strategy followed, for example, by CEO [23]. A drawback of many of these approaches is that they do not take into account conservative substitutions (each amino acid is just considered a different 'symbol' in the mutual information calculation). Small MSAs (low number of sequences) or those with uneven representations of the family diversity can also negatively influence these methods. In general, they have the advantage that they can work with (indeed they are benefited from) large MSAs with hundreds or even thousands of sequences.

A family of methodologies for detecting SDPs uses a vectorial representation of the MSA in which each protein is represented as a vector in a high-dimensional space, defined based on its amino-acid sequence. This vectorial space can be reduced to a low-dimensional one, preserving most of its information using standard statistical techniques such as Principal Component Analysis. In this space, vectors representing similar proteins (subgroups, subfamilies) cluster together (e.g. Figure 2). The ability of these methods to detect SDPs resides in a similar vectorial treatment for the individual residues, which generates an equivalent space where the residue clusters colocate with those of the families they are SDPs for. The SequenceSpace program was the first in implementing this approach [24]. A more recent representative of this type of methods is S3Det [25] (Figure 2). These methods do not take into account conservative substitutions either. The residue and protein 'spaces' generated by these methods have been shown to be rich sources of evolutionary and functional information when interactively manipulated/inspected by an expert.

Another strategy for locating SDPs is to look for positions whose pattern of change ('mutational behavior') resembles that of the whole MSA, as that is the expected behavior for these subfamily-dependent conserved positions: i.e. in an SDP position, similar amino acids correspond to globally similar proteins, and the other way around. As a representative of this strategy, in Xdet [26, 27] the pattern of change of a position is represented by a matrix containing the similarities for all pairs of amino acids at that position, defined according to a standard substitution matrix. The 'mutational behavior' of the whole MSA is represented by an equivalent matrix containing the global similarities for the corresponding pairs of proteins. The positions with matrices most similar to that of the whole alignment are selected as the predicted SDPs. These approaches do take into account conservative substitutions. As a drawback, they are not suitable for large MSAs because the matrices they use for representing the data grow quadratically with the number of sequences.

A typical extension commonly applied to all these methodologies consists in incorporating information on the 3D structure of the protein if available: generally filtering the initial set of SDPs to those in the surface of the protein and/or clustered together in 3D, as these are the expected structural characteristics of a region involved in functional specificity. Nevertheless, there are also methods that make explicit use of this 3D information, instead of using it as a posteriori filter. For example, Landgraf *et al.* [28] use 3D information to define surface patches (as sets of neighboring surface residues) and compare their conservation patterns with those of the whole family. Some methods use 3D information to create structural alignments of distant homologous, which cannot be obtained using solely sequence data. These alignments covering long evolutionary distances are then used for locating the functional subgroups and SDPs [29, 30].

Finally, we can define another group with the 'ensemble' approaches. These combine either different SDP detection methods or SDPs with other function-related patterns (coevolution, conservation, energy-based criteria), e.g. [31, 32].

Note that most of these approaches do not take as input an explicit classification of the proteins in subgroups (Figure 1), but they infer, in one way or another, the subfamily composition from the sequence relationships in the same MSA. In some contexts they are called 'unsupervised methods', to differentiate them from the 'supervised' ones, which, apart from the MSA, take a subfamily classification as additional input (which may or may not coincide with that inferred from the sequence relationships represented by the MSA) [27]. It is easy to see that most of the unsupervised methods described can be turned into supervised: e.g. forcing a given grouping of the sequences instead of obtaining it from partitions of a phylogenetic tree, forcing the sequence clusters in the sequence space or imposing an external protein similarity matrix instead of extracting it from the MSA.

We have described some representative methods for each of the main approaches. The aim of this review is not to discuss all available methods. For other more exhaustive reviews focused on methods see [31–35]. Some of these also include benchmarks and comparisons of the different methodologies. We also refer to [4, 18] for a more extensive discussion of SDP detection methods and the relationship of SDPs with other mutational patterns observed in MSAs. Finally, a good practical tutorial on how to generate MSAs and extract conservation patterns from them, including SDPs, can be found in [34].

While most of the methods described above are available as stand-alone programs or through Web interfaces (e.g. [36]), some packages have been specifically developed to facilitate the daily work with SDPs and other MSA conservation patterns related to function. JDet [37] is a graphical interactive multiplatform open-source package for the interactive calculation and visualization of function-related conservation patterns in MSAs and structures (Figure 2). The user has to provide a MSA as only input for the program. Two methods for the detection of SDPs are included in the package (Xdet and S3Det, commented above), and others can be incorporated as plug-ins or their predictions imported. One of the methods, S3Det, also reports the subfamilies automatically found in the MSA (colors in Figure 2). JDet also contains some useful features for working with MSAs and SDPs, such as a generator of sequence logos, the possibility of mapping the predicted SDPs in 3D structures and some MSA editing capabilities. The goal of this software is to make these approaches closer to the experimental researches so that they can apply them to their protein(s) of interest.

Conclusion

SDPs complement fully conserved positions as predictors of functionality. Identifying positions responsible for functional specificity is crucial for understanding and modifying protein function. From the examples described in this review, it is clear that such analysis allows designing proteins with subtle functional variations or interchanged specificities, opening interesting biotechnological possibilities. Although new methods continue to be developed, we can say that the existing approaches for detecting SDPs are mature enough for being incorporated into the toolboxes of researches in the biotechnological and biomedical areas. This 'maturity' includes professional and easy-to-use graphical interfaces, which allow their usage by non-bioinformaticians. Also contributing to the widespread use of these methodologies is the continuous increase in the number of known protein sequences that constitute their basic input. As shown by the examples discussed here, they are being successfully applied to protein families of biomedical and biotechnological interest, and we expect this trend to continue in the future. The objective of this review is to make the potential users of these approaches aware of their possibilities and the landscape of methods available. These users can now go to more specific reviews and benchmarks focused on the methods (e.g. [31–35]) to look for that more suitable for their specific need.

Key Points

- Specificity-determining positions (SDPs) complement the classic fully conserved positions as predictors of protein functionally important sites. SDPs point to protein regions involved in functional specificity.

- The performance of the methods for detecting these positions is continuously increasing owing to methodological improvements as well as the ever-increasing repertory of known protein sequences.
- These approaches are mature enough for being incorporated into the toolboxes of molecular biologists and are being applied to protein families of biotechnological and medical interest so as to get insight into the atomic basis of their functional specificities, which is essential to design mutants with interchanged or new specificities.

Acknowledgements

We thank the members of the Computational Systems Biology Group (CNB-CSIC), and Antonio Rausell (SIB), David Juan and Alfonso Valencia (CNIO) for interesting discussion. We also thank one of the anonymous reviewers of a previous version of this manuscript for convincing us to send it to a broader audience journal.

Funding

This work was partially supported by the Spanish Ministry for Economy and Competitiveness [project BIO2010-22109].

References

1. Shendure J, Hi H. Next-generation DNA sequencing. *Nat Biotechnol* 2008;**26**:1135–45.
2. van Dijk EL, Auger H, Jaszczyszyn Y, *et al*. Ten years of next-generation sequencing technology. *Trends Genet* 2014; **30**:418–26.
3. Pazos F, Sanchez-Pulido L. *Protein Superfamilies*. *eLS*. Chichester: John Wiley & Sons, Ltd, 2014, DOI: 10.1002/ 9780470015902.a9780470025587.
4. Pietrosemoli N, Lopez D, Segura-Cabrera A, *et al*. Computational prediction of important regions in protein sequences. *IEEE Sign Proc Mag* 2012;**29**:143–7.
5. Valdar WS. Scoring residue conservation. *Proteins* 2002;**48**: 227–41.
6. Valencia A, Sander, C. The ras superfamily. In: M Zerial, LA Huber, J Tooze (eds). *Guidebook to the Small GTPases*. Oxford, New York, Tokyo: Oxford University Press, 1995, 12–19.
7. Wennerberg K, Rossman KL, Der CJ. The Ras superfamily at a glance. *J Cell Sci* 2005;**118**:843–846.
8. Bauer B, Mirey G, Vetter IR, *et al*. Effector recognition by the small GTP-binding proteins Ras and Ral. *J Biol Chem* 1999;**274**: 17763–70.
9. Morillas M, Gomez-Puertas P, Bentebibel A, *et al*. Identification of conserved amino acid residues in rat liver carnitine palmitoyltransferase I critical for malonyl-CoA inhibition. Mutation of methionine 593 abolishes malonyl-CoA inhibition. *J Biol Chem* 2003;**278**:9058–63. Epub 2002 Dec 9023.
10. Cordente AG, Lopez-Viñas E, Vazquez MI, *et al*. Redesign of carnitine acetyltransferase specificity by protein engineering. Modification of methionine564 broadens the specificity to longer acyl-CoAs as substrates. *J Biol Chem* 2004;**279**: 33899–908.
11. Rodriguez GJ, Yao R, Lichtarge O, *et al*. Evolution-guided discovery and recoding of allosteric pathway specificity determinants in psychoactive bioamine receptors. *Proc Natl Acad Sci USA* 2010;**107**:7787–92.
12. Ratnikov BI, Cieplak P, Gramatikoff K, *et al*. Basis for substrate recognition and distinction by matrix metalloproteinases. *Proc Natl Acad Sci USA* 2014;**111**:E4148–55.
13. Nagase H, Visse R, Murphy G. Structure and function of matrix metalloproteinases and TIMPs. *Cardiovasc Res* 2006;**69**(3): 562–73.
14. Espaillat A, Carrasco-Lopez C, Bernardo-Garcia N, *et al*. Structural basis for the broad specificity of a new family of amino-acid racemases. *Acta Crystallographica D* 2014;**D14**: 79–90.
15. Chagoyen M, Carrascosa JL, Pazos F, *et al*. Molecular determinants of the ATP hydrolysis asymmetry of the CCT chaperonin complex. *Proteins* 2014;**82**:703–7.
16. Brill S, Sade-Falk O, Elbaz-Alon Y, *et al*. Specificity determinants in small multidrug transporters. *J Mol Biol* 2015;**427**:468–77.
17. Chevalier F, Nieminen K, Sanchez-Ferrero JC, *et al*. Strigolactone promotes degradation of DWARF14, an a/b hydrolase essential for strigolactone signaling in arabidopsis. *Plant Cell Online* 2014;**26**:1134–50.
18. Juan D, Pazos F, Valencia A. Emerging methods in protein co-evolution. *Nat Rev Genet* 2013;**14**:249–61.
19. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol* 1996;**257**:342–58.
20. Mihalek I, Res I, Lichtarge O. A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J Mol Biol* 2004;**336**:1265–82.
21. Hannenhalli SS, Russell RB. Analysis and prediction of functional sub-types from protein sequence alignments. *J Mol Biol* 2000;**303**:61–76.
22. Mirny LA, Gelfand MS. Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors. *J Mol Biol* 2002;**321**:7–20.
23. Reva B, Antipin Y, Sander C. Determinants of protein function revealed by combinatorial entropy optimization. *Genome Biol* 2007;**8**:R232.
24. Casari G, Sander C, Valencia A. A method to predict functional residues in proteins. *Nat Struct Biol* 1995;**2**:171–8.
25. Rausell A, Juan D, Pazos F, *et al*. Protein interactions and ligand binding: from protein subfamilies to functional specificity. *Proc Natl Acad Sci USA* 2010;**107**:1995–2000.
26. del Sol Mesa A, Pazos F, Valencia A. Automatic methods for predicting functionally important residues. *J Mol Biol* 2003; **326**:1289–302.
27. Pazos F, Rausell A, Valencia A. Phylogeny-independent detection of functional residues. *Bioinformatics* 2006;**22**: 1440–8.
28. Landgraf R, Xenarios I, Eisenberg D. Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J Mol Biol* 2001;**307**:1487–502.
29. Pazos F, Sternberg MJE. Automated prediction of protein function and detection of functional sites from structure. *Proc Natl Acad Sci USA* 2004;**101**:14754–9.
30. de Melo-Minardi RC, Bastard K, Artiguenave F. Identification of subfamily-specific sites based on active sites modeling and clustering. *Bioinformatics* 2010;**26**:3075–82.
31. Chakrabarti S, Panchenko AR. Ensemble approach to predict specificity determinants: benchmarking and validation. *BMC Bioinformatics* 2008;**10**:207.
32. Chakraborty A, Chakrabarti S. A survey on prediction of specificity-determining sites in proteins. *Brief Bioinform* 2015;**16**:71–88.
33. Capra JA, Singh M. Characterization and prediction of residues determining protein functional specificity. *Bioinformatics* 2008;**24**:1473–80.

34. Benitez-Paez A, Cardenas-Brito S, Gutierrez AJ. A practical guide for the computational selection of residues to be experimentally characterized in protein families. *Brief Bioinform* 2011;**13**:329–36.
35. Teppa E, Wilkins AD, Nielsen M, *et al*. Disentangling evolutionary signals: conservation, specificity determining positions and coevolution. Implication for catalytic residue prediction. *BMC Bioinformatics* 2012;**13**:235.
36. Carro A, Tress M, de Juan D *et al*. TreeDet: a web server to explore sequence space. *Nucleic Acids Res* 2006;**34**: W110–15.
37. Muth T, García-Martín JA, Rausell A, *et al*. JDet: Interactive calculation and visualization of function-related conservation patterns in multiple sequence alignments and estructures. *Bioinformatics* 2012;**28**:584–6.