COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

Volume No: 2, Issue: 3, September 2012, e201209007, http://dx.doi.org/10.5936/csbj.201209007

CSBJ

Computational enzyme design approaches with significant biological outcomes: progress and challenges

Xiaoman Li [a], Ziding Zhang [b,*], Jiangning Song [a,c,*]

Abstract: Enzymes are powerful biocatalysts, however, so far there is still a large gap between the number of enzyme-based practical applications and that of naturally occurring enzymes. Multiple experimental approaches have been applied to generate nearly all possible mutations of target enzymes, allowing the identification of desirable variants with improved properties to meet the practical needs. Meanwhile, an increasing number of computational methods have been developed to assist in the modification of enzymes during the past few decades. With the development of bioinformatic algorithms, computational approaches are now able to provide more precise guidance for enzyme engineering and make it more efficient and less laborious. In this review, we summarize the recent advances of method development with significant biological outcomes to provide important insights into successful computational protein designs. We also discuss the limitations and challenges of existing methods and the future directions that should improve them.

Mini Review Article

Introduction

Numerous enzymes have been widely used in biotechnology, pharmaceutical and industrial processes. As biocatalysts are able to accelerate the reaction speed by a factor up to 10^{17} even in mild environments [1], researchers are keen to make certain enzymes applicable in academic, industrial and commercial fields, which has resulted in rapid progress of enzyme engineering in recent years. In particular, great efforts have been made to improve the activity, stability and substrate specificity of the enzymes and design novel catalytic activity. In order to facilitate the modification of target enzymes, a variety of methodologies have been developed. They can be roughly divided into two contrasting categories: rational design and directed evolution [2].

Rational design, the earliest approach applied to the modification of enzymes [3-5], requires the availability of detailed structural information and catalytic mechanism of the targets. Computational tools have been developed to deal with a large number of data produced in rational enzyme design. In the meanwhile, such development leads to the emergence of "*de novo* computational design" approach [6], which commonly refers to the generation of novel protein scaffolds or enzymatic activity. Limited but exciting goals have been achieved in this field [7-9], making *de novo* computational design a promising approach in enzyme engineering. As another common methodology, directed evolution, was only applied to improve desired properties of enzymes recently [10, 11], but it has quickly become a powerful and popular tool in enzyme engineering [12]. Nevertheless, the bottleneck of directed evolution lies in the development of an efficient high-throughput screening technology, despite that there are quite a few successful examples that used directed evolution to modify important commercial enzymes [13-16]. Consequently, the combined approaches involving rational or *de novo* design with directed evolution may offer significant advantages over individual approaches [8, 17].

In this mini-review, we highlight the strengths of a number of effective computational methodologies/tools that can assist in the rational and *de novo* enzyme design (see Figure 1). Successful examples, especially those concerning improvement of enzymatic activity and stability, which are the most important properties from a practical perspective, are discussed in the following respective sections.

Rational design strategies and tools

The success of rational design depends on our in-depth knowledge about sequence and structure features of target proteins. A popular strategy to identify functionally related residues of unknown targets is the use of sequence features. Analysis of these features can provide enough information about evolutionary relationship, functional sites, correlated mutations and so on. The most useful tools for extracting sequence information are multiple sequence alignment (MSA) and coevolutionary analysis [18], while the latter sometimes requires structural information. As a matter of fact, structure-based design is no doubt more efficient to locate key residues, because the execution of the protein function is directly linked with the maintenance of the 3D structure in functionally related regions. Structure-based rational design can benefit considerably from the rapidly growing number of solved protein structures, however, these account for only a small portion of naturally occurring proteins. To make a better use of structural

[a]*National Engineering Laboratory for Industrial Enzymes and Key Laboratory of Systems Microbial Biotechnology, Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin, Tianjin 300308, China*
[b]*State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University, Beijing 100193, China*
[c]*Department of Biochemistry and Molecular Biology and ARC Centre of Excellence in Structural and Functional Microbial Genomics, Monash University, Melbourne, VIC 3800, Australia*

* Corresponding author.
E-mail address: zidingzhang@cau.edu.cn (Ziding Zhang)
Jiangning.Song@monash.edu (Jiangning Song)

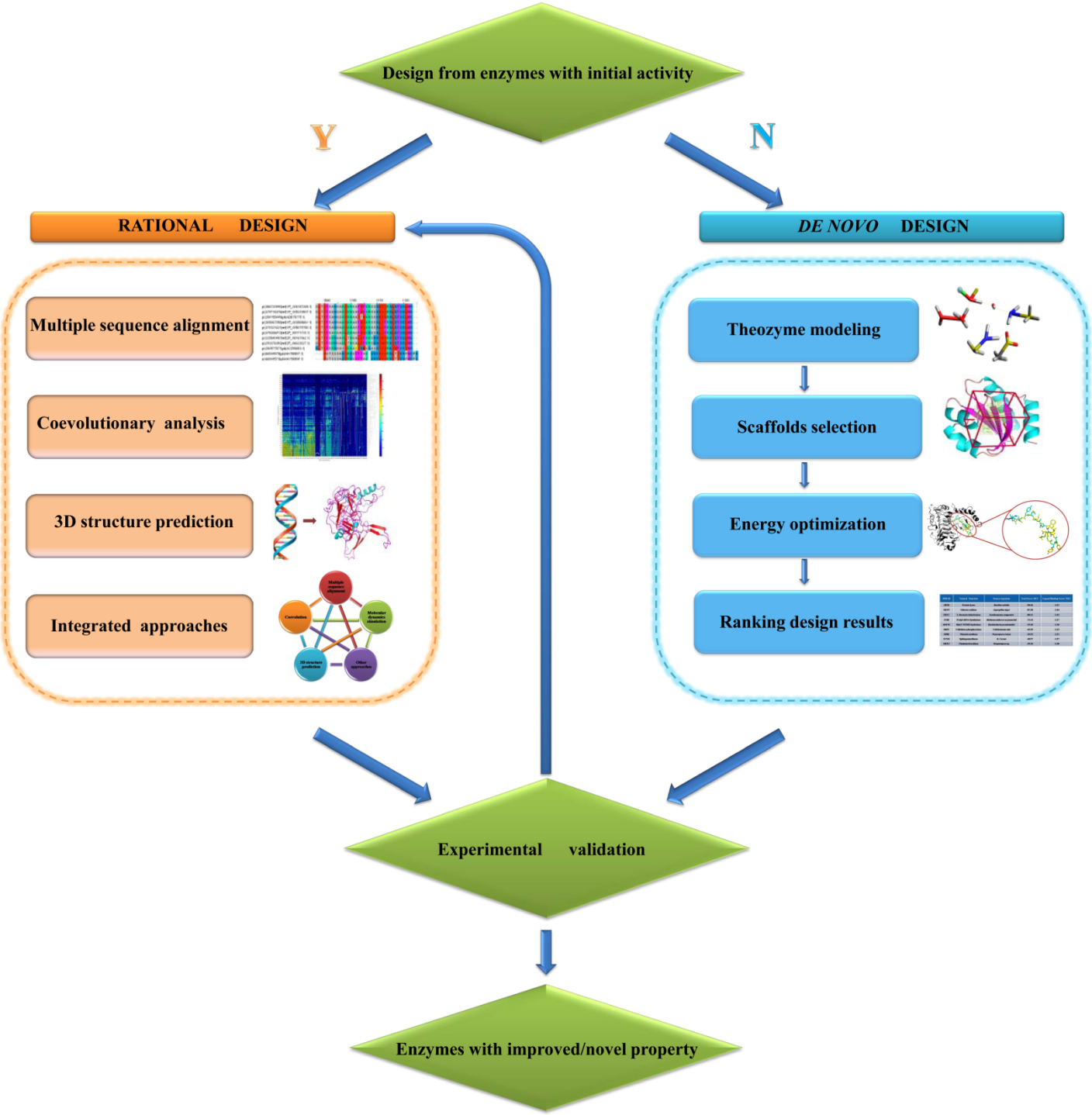

Figure 1. Strategies of rational and *de novo* enzyme design

information, 3D structure prediction or analysis tools are extremely important and greatly desired. Fortunately, a variety of computational methodologies/tools have been available to facilitate processing and data analysis, which have significantly contributed to the progress of rational enzyme design. Among them, several noteworthy tools are discussed below.

Multiple sequence alignment (MSA)

Protein primary sequence provides the most direct and readily available information for rational design, because important clues for potential mutation sites can be extracted from the amino acid sequence in cases where structural information is not available. For example, Ni *et al.* investigated the activity-related mutations in the wild type of endo-b-1,4-glucanase (RsEG) of *Reticulitermes speratus* via sequence comparison with other cellulases from different sources [19], as well as a RsEG mutant obtained from directed evolution. As a result, they obtained a higher activity and higher expression level of the RsEG mutant. Their analysis identified three single mutants that contributed to a higher enzyme activity, and four residues predicted to be located in the catalytic center by MSA analysis were also experimentally verified. In fact, sequence comparison tends to be more reliable when a reasonable number of homologous sequences are

available. High-throughput sequencing techniques have produced larger amounts of data than before. To deal with such data, a variety of MSA methods have been developed in the past two decades [20-22] and have a wide range of applications in modern molecular biology. For rational enzyme design, the construction and analysis of MSA are usually required in the identification of functional-related residues, specificity-determining positions, homology modelling and protein function prediction [22].

Using progressive alignment algorithm [23], a classical MSA method called ClustalW has been widely exploited in various research fields [24-26], and it can generally yield a better performance for highly homologous sequences [27]. For instance, Ehren and co-workers used ClustalW to construct an MSA of 100 homologues of prolyl endopeptidase (PEP) from *Sphingomonas capsulate*, and proposed a list of 30 potentially beneficial mutations based on the generated MSA [28]. A mutagenesis library with limited members was then established, facilitating the selection step and in-depth investigation of each variant. After two rounds of mutagenesis, mutants with enhanced activity and significantly raised resistance to pepsin digestion were identified. In another application of ClustalW, Gumpena *et al.* investigated different proteins from the same gluzincin family. They found that salt bridges that execute similar functions were formed by different residue pairs, and that these salt bridges were not interchangeable, indicating divergent microenvironments around active sites [29]. At present, both ClustalW and its new version ClustalOmega whose accuracy is not influenced by the size of sequences [30], are freely available to the community.

In addition to ClustalW, there are also alternative MSA tools, such as T-Coffee[31], Mafft[32] and Muscle [33], which offer a significantly improved alignment quality with, in some cases, reduced CPU time [34]. Among these, Mafft has been found to be able to provide a consistently better performance in terms of the calculation speed, high quality score with high-throughput data, and high accuracy with very divergent blocks, when evaluated on different benchmarks [35, 36]. Mafft explores two novel methods to enhance its accuracy and scalability [32], which include a fast Fourier transform algorithm that allows rapid identification of homologous regions, and a simplified scoring system designed for CPU time reduction and accuracy improvement of alignments in the case of less homologous sequences. Another iterative refinement technique is also used in Mafft to correct the errors introduced by the progressive alignment [22, 32]. The first version of Mafft was well characterized by a comparable accuracy but shorter CPU time in contrast to ClustalW and T-Coffee, and has been continuously improved in the past ten years [37-39]. The latest version of Mafft is 6.903, which can be run on Mac OS X, Linux, and Windows. Regarding the application of Mafft in protein design, Michel *et al.* compared members of the polysaccharide lyase family 6 with the chondroitin B lyase from *Pedobacter heparinus* [40]. Conserved residues that interact with Ca^{2+} ion were located precisely from the primary sequence, confirming that the chondroitin B lyase has a calcium-dependent catalytic mechanism. MSA analysis was also validated by the X-ray structure and site-directed mutagenesis. In the follow-up enzyme engineering step, the redesign of such function-related residues can be avoided in advance. Maita *et al.* also employed Mafft to perform an MSA analysis of oligosaccharyl transferases (OSTs) from different microbial domains [41]. After inclusion of a considerable number of distantly related sequences, Mafft yielded a satisfying performance and facilitated the identification of three different kinds of catalytic centers. Furthermore, they also found that two distantly related OSTs share a higher structural similarity than sequence similarity. These results indicate that the application of additional information in MSAs, such as sequence homologs and structural information, can improve the MSA quality [20].

In addition to the improvement of computational algorithms, there is another trend that involves a combination of several MSA methods based on the same set of sequences. The work on 3-deoxy-D-manno-octulosonate 8-phosphatesynthases (KDO8PS) by Ackerman *et al.* provided a good example [42]. In that work, Mafft, T-Coffee and Muscle programs were used individually for curating the MSAs of all known KDO8PS, with the results further integrated using T-Coffee. Seven pairs of coevolved residues were identified, and their contribution to protein stability was examined. Interestingly, one mutation in one coevolving residue pair that resulted in a slight decrease in protein stability could be compensated by another mutation in the same pair to maximize the stability of the protein. These results highlight that an important property, "coevolution", extracted from a curated MSA of protein sequences, can provide a meaningful research direction for rational enzyme design.

Coevolutionary analysis

Coevolution (also known as covariation, correlated mutation or co-substitution) refers to "reciprocal evolutionary change in evolutionarily interacting loci" [43], and occurs at many levels in biology [44-46]. In this review, only the correlated mutations between amino acids within a protein are discussed. Coevolutionary analysis methods have a number of important applications in the prediction of protein structure [47, 48], identification of functional sites [49-51] and candidate design sites [52, 53]. The identified coevolving residues have been experimentally validated in some studies [54, 55], implying the potential application of coevolutionary analysis in rational enzyme design.

In the past few decades, a number of coevolutionary analysis algorithms have been developed [56]. These methods share a common procedure of three steps: MSA construction, coevolutionary measure calculation and experimental validation. Most coevolutionary analyses start with the construction of an MSA of the query protein. Although certain automatic software can be applied (see Table 1), manual refinement, including filtering of sequences with large gaps, low homology or wrong annotation, is often required to ensure a high-quality MSA [57]. The second step is to calculate coevolutionary measures, which can be done by using different correlated mutation algorithms, followed by statistical significance tests and analyses to extract significant coevolution values, eliminate background noise [58] and evaluate the performance and robustness of the coevolution measures [59]. Finally, "wet" experiments need to be performed to validate the obtained coevolutionary results.

For experimental scientists, coevolutionary webservers seem to be more straightforward, attractive and practical. Up to now, several online tools have been made publicly available [56, 60]. However, how to choose an optimal scoring function of coevolutionary measures in the second step remains to be a critical factor that will determine the quality of coevolutionary analysis. To address this, Fodor *et al.* [61] assessed the performance of four different methods in detecting coevolutionary site, namely Statistical Coupling Analysis (SCA) [62], Observed Minus Expected Squared (OMES) [63], McLachlan Based Substitution correlation (McBASC) [64] and Mutual Information (MI) [57]. In their research, OMES and McBASC were found to outperform the other two algorithms in favoring poorly conserved residue pairs and decreasing sensitivity to background conservation, and were of considerable similarity in sensitivity to background noise. The OMES-based programs, OMES-KASS [63] and Fodor package [61], which were more recently developed, have been applied to perform reliable coevolutionary

Table 1. Summary of useful computational programs in rational design referred in this review.

Programs	Application	URL address	Operating system	Ref.
Rational design programs				
ClustalW	Multiple sequence alignment	http://www.clustal.org/clustal2/	Windows, Linux, MacOS	[27, 122]
ClustalOmega		http://www.clustal.org/omega/	Windows, Linux, MacOS	[30]
Mafft		http://mafft.cbrc.jp/alignment/software/	Windows, Linux, MacOS	[32, 37, 39]
T-Coffee		http://www.tcoffee.org/Projects/tcoffee/	Linux, MacOS	[123]
Muscle		http://www.drive5.com/muscle/	Windows, Linux, MacOS	[33]
Integrated system	Coevolutionary analysis	http://coevolution.gersteinlab.org/coevolution/	Windows, Linux, MacOS	[60]
OMES-KASS		http://bip.weizmann.ac.il/correlated_mutations/	Linux	[63]
Fodor package		http://www.afodor.net/	Windows, Linux, MacOS	[61]
Swiss-Model	3D structure prediction	http://swissmodel.expasy.org/	-	[124, 125]
HHpred2		http://toolkit.tuebingen.mpg.de/hhpred	-	[83]
I-TASSER		http://zhanglab.ccmb.med.umich.edu/I-TASSER/	Linux	[84, 126]
FoldX	Protein stability prediction	http://foldx.crg.es/	Windows, Linux, MacOS	[96, 127]
PopMuSiC		http://babylone.ulb.ac.be/popmusic	-	[94, 97, 128]
I-Mutant3.0		http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant3.0/I-Mutant3.0.cgi	-	[129]
DMutant		http://sparks.informatics.iupui.edu/hzhou/mutation.html	-	[130]
***De novo* design programs**				
RosettaMatch	Scaffold search	-	-	[108]
RosettaDesign	Protein design for low free energy sequences	http://rosettadesign.med.unc.edu/	Linux	[109]
ORBIT	Optimal sequences search for given folds	-	-	[118]

analysis [65-67]. In addition, Yip *et al.* developed an integrated online program by embedding several coevolutionary algorithms into one system instead of using a single algorithm only. These algorithms include SCA, MI, Explicit Likelihood of Subset Variation (ELSC) [68] and correlation-based methods [64, 69], making this system a convenient comparative analysis tool of different coevolutionary methods. The integrated system also provides an MSA preprocessing option to further improve its performance. In addition, users can also choose to treat the gaps in the MSA as noise or as an additional 21st residue, based on the observation that gaps might contain important coevolutionary information [60].

Despite the functional significance, how to combine coevolutionary analysis with rational enzyme design remains a challenging issue. In 2011, Zeng and colleagues applied SCA to analyse the sequences of the regulatory domains of the aspartokinase (AK) family to characterize the allosteric interaction network [53] and integrated such information with rational enzyme design. AK is the central enzyme in the biosynthesis of aspartate family amino acids, and the allosteric inhibition of AK by end-products obstructs the production of related amino acids in *Corynebacterium glutamicum* [70]. As a result, their coevolutionary analysis of 500 sequences from the AK family identified 25 highly correlated positions, in which 14 sites were mutated to construct AK mutants of *C.glutamicum*. All the mutants showed resistance to allosteric inhibition to different extents, suggesting that the choice of target mutations was largely successful. In this study, a major strategy was to select residues that had the potential to interrupt allosteric interaction, whereas in researches that aim to modify other properties of enzymes, amino acidsites that regulate the target property can probably be selected as candidates according to expert knowledge or structural analysis. There were two general rules to mutate the wild-type amino acids at the selected sites: (i) mutating the wild-type amino acids to those with less usage frequency at the corresponding positions; (ii) or substituting the wild-type amino acids by those with different chemical properties with the purpose of making more obvious changes in terms of the target properties [53]. In another work of Chen and co-workers, AK3 from *Escherichia coli* was investigated via an integrative analysis of coevolution and molecular dynamics (MD) [71]. The SCA-based coevolutionary analysis of 340 protein sequences with 424 positions was combined with the 10 nanosecond (ns) MD simulation of AK3

with/without lysine as an effector molecule. 30 top ranked positions were accordingly selected, most of which were reported as potential targets for point mutations in other studies using random mutagenesis. The site-directed mutations of the remaining positions not found by random mutagenesis, however, led to significant deregulation of allosteric inhibition by effectors. Although both coevolutionary analysis and MD simulation are complicated, usually requiring iterative procedures prior to the result generation, they have better efficiencies than traditional experimental approaches like random mutagenesis. In the case of AK3, its computational design can be "grafted" into another AK of the same family even with lower sequence identity, making it more efficient and appealing.

Protein 3D structure prediction

There are an increasing number of proteins with high-resolution solved 3D structures, greatly facilitating the rational and computational protein design. Numerous previous successes have shown that when 3D structural information is available, protein design can be much more precise and accurate [18, 72, 73]. It is apparent that the knowledge of 3D structure of the target enzyme is a prerequisite and foundation for structure-based design. Although only a small portion of proteins have authentic crystal structures, those with unknown structure information can be reliably modeled via protein 3D structure prediction software, provided that there is a known structure of one or several homologous proteins to the target protein [74, 75].

According to the availability of template structures, protein 3D structure prediction can be generally divided into two categories: homology modelling and *ab initio* modelling. The former refers to the construction of an atomic-resolution model of a protein from its primary sequence using the experimentally solved 3D structure of a homologous protein as the "template", while the latter is called "free modelling" or "*de novo* modelling" in some cases, referring to 3D structure prediction generated from scratch when structural analogs are not available or detectable. The majority of methods used in homology modelling can be further grouped into two types: comparative modelling (CM) [76] and threading [77]. The root mean square deviation (rmsd) of a CM constructed model from the structure obtained from experiments can usually achieve 1–2 Å when a highly homologous (>30% sequence identity) template is employed. Models with such accuracy can compete with the low-resolution X-ray or medium-resolution NMR structures [78]. In contrast, the threading approach usually has a remarkable performance when dealing with target protein modelling using relatively distant templates, and the corresponding rmsd is 2-6Å [79] with most errors occurring in loops. *Ab initio* modelling, however, continues to be the most challenging topic in protein 3D structure prediction. Although there has been an exciting progress in modelling small proteins, no substantial progress has been achieved in *de novo* structure prediction of proteins with more than 150 residues [80]. In view of this, we mainly focus on the homology modelling methods in this mini-review.

According to the initial plan of protein Structure Initiative (PSI), proteins within 90% of the domain families can be modeled by CM at its completeness [81]. As a consequence of this project, homology modelling is becoming increasingly important. Nowadays, a handful of academic-free servers for template-based protein structure prediction are available without any restrictions, resulting in a confusion about which tool should be used for solving different tasks. A popular criterion to assess the 3D structure prediction quality is the Critical Assessment of Structure Prediction (CASP) which has been carried out each two years since 1994 [76]. In the latest competition, CASP9 in 2010 [82], 176 groups took part in the homology modeling which is the most relevant category for biological applications. According to the results of the assessment, a group of six methods have outperformed noticeably the rest ones in the "server" category [82], among which HHpreß [83] and Zhang-Server (namely I-TASSER) [84, 85] were assessed as the best.

However, no matter how significantly an algorithm has been improved, the modelling quality greatly relies on the sequence homology between the template and the target. The prediction procedure can be further simplified and become straightforward when a closely related template is available. Besides, meta-server, which produces a combined prediction using results of other automatic servers, has proved to outperform most individual ones [86]. Due to page limitation, only the popular automated webservers that suit protein design purposes are reviewed in this section. Swiss-Model [87], an automated CM server, is regarded as the most widely used online tool in protein 3D structure prediction. CM, as described above, is the only methodology that can reliably model a 3D structure using amino acid sequence alone [88]. By submitting an amino acid sequence or its UniProtID, users start the modelling procedure with or without providing a template protein. Swiss-Model server can automatically select several suitable templates from a refined library derived from the Protein Data Bank (PDB), and then a structural alignment between the target and the template is generated and improved for the sake of modelling [87]. The mapping of the residue correspondence between the target and the template begins at this step, followed by model building. In the Swiss-Model server, three building modes can be selected before the submission: "automated", "alignment" and "project"; it is recommended to choose options according to the similarity between the target and templates [89]. "Automated" is for higher similarity of >60%, "project" for that below 20%, and "alignment" otherwise. The energy minimization of the built models by the GROMOS96 force field is the final step. Efforts have been made to improve the modelling quality of Swiss-Model since it was developed. Numerous examples have been provided in literature, and some representatives are discussed here. The Kir2 channels are a kind of potassium selective channels [90]. A pH sensitive member Kir2.3 was aligned with all the Kir2 channel proteins, and histidine 117 (H117) located close to the putative selectivity filter was identified to contribute to pH sensitive phenotype [91]. However, contradictory results were obtained by directed mutagenesis experiments, suggesting that there were other factors related to the pH effect. The observation that the ability of Zn^{2+} to bind cysteines/histidines could inhibit the pH effect indicated that a cysteine within atomic distance to H117 might interact to exert this functional effect. Consequently, the 3D structure of Kir2.3 was created by Swiss-Model using distant templates in order to narrow down the range and locate the target cysteine. The rational design of candidate sites was implemented by site-directed mutagenesis, and C141 was found to interact with H117 to exert an influence on pH sensitivity. In another example, Choi and colleagues carried out homology modelling-based rational design of an epoxide hydrolase (EH) in a marine fish, *Mugil cephalus* [92]. The 3D structure of EH from a fungus, *Aspergillus niger,* was selected as the template by Swiss-Model for 3D structure prediction of *M.cephalus* EH. The active sites of the predicted structure were then superimposed on the template and indicated that the spatial orientation of D199 in the target EH was different from its counterpart in the template. Attempts to modify D199 into a proper orientation were also made to redesign the surrounding residues so that they could have direct or indirect interactions with D199. To achieve this, F193 and Y194 were chosen, and the 3D structures of various mutants of these two residues were constructed by Swiss-Model instead of "wet" experiments. Analysis of the corresponding 3D structures, particularly

the activity sites, revealed that D199 had the right orientation in the variants F193Y and Y194M. Site-specific mutagenesis confirmed that the F193Y mutant indeed improved the catalytic activity and decreased the reaction time. It is worth noting that the reliability of Swiss-Model prediction was validated in a situation where a distant template was used, providing a good example of freeing researchers from laborious experiments by entirely resorting to the Swiss-Model computational tool.

A closely related issue is protein stability design. In a recent work on glycerol dehydratase (GDHt) [93], prediction of protein stability was realized by a computational program called PoPMuSiC [94]. The selection of point mutation residues mainly depended on the prediction result. The performance of such tool requires the 3D structure of the target protein. Accordingly, homology modelling of the target GDHt was first conducted by the Swiss-Model server based on the template retrieved from PDB (ID: 1IWP). Two mutations that were predicted to be the most stable were selected and mutated by single point mutation. The 3D models of the two mutants were built again using Swiss-Model. An enhanced hydrogen bond interaction between the mutated positions and the surrounding residues accounted for the improved stability, which was validated by experiments. We conclude from a large number of examples including those discussed above that 3D structure prediction provides not only direct evidence for rational protein design, but also essential assistance for structure-based enzyme redesign. Since less than 1% proteins have solved 3D structures, studies on the stability and other important properties of most target proteins have to rely on the predicted structural information.

Unfortunately, there are no generally applicable rules for enzyme activity enhancement, due to the variance in catalytic mechanisms of different types of enzymes. Therefore, many efforts have also been made to improve other important properties of enzyme catalysts, for example, protein stability, a critical property of an enzyme catalyst that is pertinent to its industrial potential. As Swiss-Model and many other predictors can produce high-quality results, a crucial step in protein stability prediction is the choice of well-performing servers. According to a recent systematic analysis of 11 online stability predictors by Khan and Vihinen [95], FoldX [96] is amongst the top ones. However, FoldX does not provide a convenient online webserver, which has limited its broad application. Another well-performing tool PoPMuSiC provides an alternative choice, which was developed in 2000 [97] and updated in 2009 [98] using more experimental data from ProTherm [99]. The most-recent version of PoPMuSiC webserver was released in 2011 [94], providing a systematic evaluation on stability changes under saturated single-site mutations at each residue position, or an appointed one for the submitted protein on the basis of its 3D structure.

De novo computational design

The ultimate test of our understanding of the mechanism of enzymatic catalysis is *de novo* computational design, which refers to creation of novel protein folds, substrate binding pockets, and catalytic activities and so on. *De novo* protein design was first conducted to create a four-helix bundle protein in 1988 [6]. Since then, various protein folds have been *de novo* designed [100]. However, only a few possessed catalytic functions. Accordingly, *de novo* computational design of naturally occurring enzymes with novel catalytic activity is considered as a grand challenge, and in recent years, great efforts in this field have been made to expand our knowledge in enzyme engineering [7-9, 101-103]. To illustrate this, in this section we discuss three distinguished design examples of enzymes that catalyze synthetic reactions.

The overwhelming performance of enzymatic catalysis over chemical catalysis is partly due to the free energy decrease of transition state (TS) via the interaction with catalytic residues [104]. Hence, the first step of *de novo* design for a given reaction is to model its theozyme which consists of TS model and catalytic groups [105] based on quantum chemical calculations [106]. How well the theozyme models correlate with their corresponding crystal structures, will have a significant influence on the ultimate designs. Dechancie *et al.* mimicked the active sites of nine distinct enzymes with quantum mechanical optimizations [107]. The rmsd of the sets of catalytic atoms was 0.64Å, suggesting that the predicted geometries were remarkably consistent with the corresponding X-ray structure. For a desired reaction, there usually exist more than one possible catalytic mechanism. As result, the 3D models of each catalytic motif for each mechanism will have to be built, and hence the degree of freedom and the orientation of different bonds in each model can vary greatly, giving rise to a great number of possible 3D active sites, which are called "theozyme library".

The search for optimal protein scaffolds that are able to fulfill a target reaction can be launched once the theozyme library has been generated. Numerous scaffolds with ligand-binding cavities and high-resolution X-ray structures are available in several public protein databases. If there are certain restrictions on potential scaffolds, for example, in cases where a thermophilic scaffold is required, the selection range could be narrowed down. However, this process depends on *de novo* design algorithms such as RosettaMatch [108] that relies on hashing techniques and pruning of the majority of potential active centers at a very high speed but very little cost. At this step, the description of TS and a set of protein scaffolds are input into RosettaMatch. Once a TS position is compatible with the geometry of catalytic sites in one scaffold and satisfies other catalytic constraints, the result will be output as a "match" [106, 108].

Because there are still substantial candidate matches after the scaffold selection, and there remain certain steric clashes between the TS position and the catalytic side chains in the matches, further optimization is necessary. In this regard, the RosettaDesign methodology [109] can be applied to improve the binding affinity to TS and the stability of the active centers by redesigning or repacking of related residues. It is suggested that users run a single task for ten times owing to astochastic sampling algorithm adopted by RosettaDesign which will probably produce 10 distinct outputs. The resulting designs are supposed to be lower energy sequences for a given scaffold with the maximized TS affinity.

After optimizing all unique matches, a next step is to select designs with optimal performance for experimental validation. Several important factors, especially the ligand binding energy feature, are often used to evaluate and rank all the designs as described in [106]. As it is unlikely that a design has the highest score for each factor, extensive examinations to assist in further selection are preferred. In addition, Kiss *et al.* found that the MD technology was the most effective procedure for predicting the catalytic potentiality of designs [110].

The same protein scaffolds can execute diverse functions, such as α/β–barrel motif, which constitutes approximately 10% of proteins that perform a wide range of catalytic reactions [111]. This indicates that the designable potentiality of certain scaffolds underlies the foundation of computational engineering of novel functions. With similar strategies, Baker's group has performed a series of pioneering studies in redesigning enzymes that catalyze retro-aldol reaction [7], Kemp elimination [8] and Diels-Alder reaction [9].The enhancement of target reactions by designed enzymes was assessed by the ratio of

Table 2. Summary of representative examples referred in this review.

Enzyme/protein	Target property	Method		Result			Ref.
		Design strategy	Bioinformatic tool	No. of mutants	Fold-improvement	Library size	
*R.speratus*endo-β-1,4-glucanase	Activity	Functional and activity-related residues identified via an MSA analysis of eight sequences	-	7	7-13	24	[19]
S.capsulata prolyl endopeptidases	Activity; stability	An MSA of 100 homologues evaluated by multiple scoring functions identified mutations	ClustalW, SeqDist, KaKs, probCons, SUB	6(1st round) 9(2nd round)	20%(activity) 200(stability)	47(1st round) 48(2nd round)	[28]
KDO8P Synthase family	Stability	Integrated analysis by MSA, ΔΔG changes calculation, MD simulation and coevolutionary analysis	Mafft, T-Coffee, Muscle, HMMER 3.0, Prime 2.1, Desmond, X-Cluster, FoldX	No experimental validation			[42]
C.glutamicum aspartokinase	Allosteric inhibition	Correlated positions were identified by coevolutionary analysis of 500 sequences	Muscle, ClustalX	1	2	14	[53]
E.coli aspartokinase	Allosteric inhibition	Integrated analysis byMD simulation and coevolutionary analysis	Modeller, AMBER, Muscle	6	5-7	18	[71]
M. cephalus epoxide hydrolase	Activity	Activity-related residues were identified by superimposition of a predicted structure and a solved structure template	Swiss-Model, RasMol, Deep-View	1	35	5	[92]
K. pneumonia glyceroldehydratase	pH stability; activity	Stability-related residues were designed based on a predicted structure	Swiss-Model, PoPMuSiC	1	2(pH stability) 2(activity)	2	[93]
Retro-aldol reaction	Activity	TS simulated by QM/MM was used for scaffold selection and followed by individual optimization and ranking	RosettaMatch, RosettaDesign	32	2×10^4	72	[7]
Kemp elimination				8	$>10^6$	59	[8]
Diels-Alder reaction				2	89M	84	[9]
E. coli thioredoxin	PNPA hydrolase	Potential active sites and surrounding active-site mutations were identified and computed	ORBIT	2	180	2	[115]
Sperm whale myoglobin	Nitric oxide reductase	Creating a non-haem Fe^{2+}-binding site based on the predicted structure overlaid with the reference structure	VMD, NAMD	1	N.A.	<10	[117]

N.A.: Not Available

the catalytic rate constant and uncatalyzed rate constant k_{cat}/k_{uncat}. In the above cases, the values of k_{cat}/k_{uncat} ranged from 10^2 to10^5 for the most active designs, indicating the effectiveness of such design strategies. *De novo* computational enzyme design provides important insights into the structure-function relationship of the enzyme and the starting points for directed evolution and rational design. Considerable experimental efforts, including development of technologies discussed in the *Rational design strategies and tools* section, were made to enhance the activities of the artificial Kemp eliminases [112-114].

While the *Rosseta*-based *de novo* design is well characterized by its own scaffold selection steps, it is worth noting that other types of *de novo* design approaches are emerging recently and have achieved an impressive success, which were also developed based on a given scaffold [101, 115-117]. Once a suitable protein scaffold is selected according to the desirable properties of the target reaction, such as thermostability, high expression level and presence of cofactor binding domain, *de novo* design approaches only need to build an activity center and a substrate/cofactor binding pocket. In this regard, Bolon and Mayo presented a representative example of a "compute and build" strategy [115]. They chose *E. coli* thioredoxin as the starting scaffold due to its favorable thermostability and expression in *E. coli*, and used the p-nitrophenyl acetate (PNPA) hydrolyzation regulated by a histidine as the target reaction. A computational algorithm called ORBIT (optimization of rotamers by iterative technique) [118], was applied to scan active sites in the starting scaffold. Two catalytic sites were identified, and mutations surrounding these catalytic sites were then introduced in order to build compatible substrate binding pockets. Two resulting designs were further experimentally validated. One design PZD2 reached a k_{cat}/k_{uncat} value of 180. In another example, *de novo* design of a functional metalloprotein, namely the nutric oxide reductase (NOR), was performed by Yeung *et al.* [117]. The goal was to build a non-haem Fe^{2+}-binding site (Fe_B) in the scaffold sperm whale myoglobin (Mb). Based on the structural information of a structural homologue with a haem-copper site, two residues L29 and F43 were mutated to histidines, which constituted the Fe_B center together with H64 and V68E. Modelling analysis using an extension of Visual Molecular Dynamics (VMD) was performed to build the designed protein Fe_BMb, suggesting the formation of the Fe_B. Subsequent crystal and experimental data confirmed the accuracy of the predicted model and an apparently increased activity. These examples discussed above highlight the importance and complementarity of these alternative *de novo* design strategies, which can be applied to similar scaffold-based studies.

Discussions

In this mini-review, we aim to provide a useful guide on the selection of the basic design methodologies and tools that are frequently employed in enzyme engineering (Table 1), and a brief summary of these aforementioned examples is depicted in Table 2. For many naturally occurring enzymes, it is often necessary to modify and design their properties in order to meet the needs of commercial or industrial applications. Bioinformatic strategies and tools, particularly those with freely accessible webservers, offer biologists tremendous help to narrow down their experimental efforts.

MSA can efficiently identify consensus, highly conserved and variable positions within a family of homologous proteins, while MSA-based coevolutionary analysis of a set of enzymes with similar functions provide critical clues about catalytic and other functionally related residues. A number of candidate sites derived from these sequence-based studies can be used to construct a mutation library, facilitating the discovery of favorable mutants with improved functional properties. On the other hand, with the increasing availability of high-quality 3D structures in the PDB, there are a growing number of structure-based approaches being developed. Because experimentally solved structures only cover a limited portion of the protein repertoire, sequence-based 3D structure prediction has become a prevalent methodology in enzyme engineering. This is important, because reliable prediction of protein structure can still provide valuable information regarding potential candidate sites whose mutations might lead to improved properties of the enzyme, even if its real structural information is not at hand. As a symbol of the engineering of the third wave of biocatalysts [119], *de novo* enzyme design has achieved a significant success in the last 20 years. Despite these advances, there are challenges for rational enzyme design. A first challenge is that there are inevitable experimental errors in "wet" experiments [120], resulting in less reliable designs based on such low-quality data. A second challenge is related to the conformational dynamic nature of the enzyme. Conformational changes of the enzyme are frequently occurring under catalytic conditions, leading to a deviation of the real orientation of residues and enzyme structure from that of the designed or modeled enzymes. A third challenge is how to select the most appropriate tool that best suits the study of a particular target enzyme, from a pool of different tools that have both pros and cons. In this mini-review, we attempt to provide a useful guide to summarize some of the popular, reliable and academic free tools. Moreover, many examples have proved that integrative strategies can usually outperform individuals. In this regard, development of meta-servers is promising for providing a better performance and reliability of computation design. A fourth challenge is that some modified catalysts still cannot meet the practical needs of large-scale applications, particularly *de novo* designed enzymes. As such, there is often a need for assistance of experimental approaches, such as directed evolution. In fact, the boundary of rational design and directed evolution has become more and more blurred in practical applications, as evidenced by a number of recent studies that involve a combination of both [5]. Therefore, improving experimental techniques, such as high-quality mutagenesis and high-throughput screening, is another related future direction.

Due to the aforementioned challenges, many attempts of computational protein design failed. However, future development of the field will be advanced by a better understanding of the underlying reasons that led to both failures and successes [121]. Recent advances in computational enzyme design have largely expedited the evolution of enzymes, and have greatly revolutionized the way of enzyme engineering. With the development of improved experimental techniques, computational enzyme design will gain a momentum and achieve significant successes in the future.

Acknowledgements

This work was supported by grants from the Hundred Talents Program of the Chinese Academy of Sciences (CAS), the National Natural Science Foundation of China (No. 61202167), the Knowledge Innovative Program of CAS (No. KSCX2-EW-G-8) and Tianjin Municipal Science & Technology Commission (No. 10ZCKFSY05600) and the National Health and Medical Research Council of Australia (NHMRC) (No. 490989). JS is an NHMRC Peter Doherty Fellow and a Recipient of the Hundred Talents Program of CAS and the Japan Society for the Promotion of Science (JSPS) Short-term Invitation Fellowship to the Bioinformatics Center, Kyoto University, Japan.

Citation
Li X, Zhang Z, Song J (2012) Computational enzyme design approaches with significant biological outcomes: progress and challenges. Computational and Structural Biotechnology Journal. 2 (3): e201209007. doi: http://dx.doi.org/10.5936/csbj.201209007

Competing Interests:
The authors have declared that no competing interests exist.

Received: 24 July 2012
Received in revised form: 27 September 2012
Accepted: 04 October 2012

References

1. Radzicka A, Wolfenden R (1995) A proficient enzyme. Science 267: 90-93.
2. Bornscheuer UT, Pohl M (2001) Improved biocatalysts by directed evolution and rational protein design. Curr Opin Chem Biol 5: 137-143.
3. Villafranca JE, Howell EE, Voet DH, Strobel MS, Ogden RC, et al. (1983) Directed mutagenesis of dihydrofolate reductase. Science 222: 782-788.
4. Craik CS, Largman C, Fletcher T, Roczniak S, Barr PJ, et al. (1985) Redesigning trypsin: alteration of substrate specificity. Science 228: 291-297.
5. Chica RA, Doucet N, Pelletier JN (2005) Semi-rational approaches to engineering enzyme activity: combining the benefits of directed evolution and rational design. Curr Opin Biotechnol 16: 378-384.
6. Regan L, DeGrado WF (1988) Characterization of a helical protein designed from first principles. Science 241: 976-978.
7. Jiang L, Althoff EA, Clemente FR, Doyle L, Rothlisberger D, et al. (2008) De novo computational design of retro-aldol enzymes. Science 319: 1387-1391.
8. Rothlisberger D, Khersonsky O, Wollacott AM, Jiang L, DeChancie J, et al. (2008) Kemp elimination catalysts by computational enzyme design. Nature 453: 190-195.
9. Siegel JB, Zanghellini A, Lovick HM, Kiss G, Lambert AR, et al. (2010) Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. Science 329: 309-313.
10. Roberts BL, Markland W, Ley AC, Kent RB, White DW, et al. (1992) Directed evolution of a protein: selection of potent neutrophil elastase inhibitors displayed on M13 fusion phage. Proc Natl Acad Sci U S A 89: 2429-2433.
11. Osuna J, Flores H, Soberon X (1994) Microbial systems and directed evolution of protein activities. Crit Rev Microbiol 20: 107-116.
12. Dalby PA (2003) Optimising enzyme function by directed evolution. Curr Opin Struct Biol 13: 500-505.
13. Asano Y, Kira I, Yokozeki K (2005) Alteration of substrate specificity of aspartase by directed evolution. Biomolecular Engineering 22: 95-101.
14. Bastian S, Rekowski MJ, Witte K, Heckmann-Pohl DM, Giffhorn F (2005) Engineering of pyranose 2-oxidase from Peniophora gigantea towards improved thermostability and catalytic efficiency. Appl Microbiol Biotechnol 67: 654-663.
15. O'Loughlin TL, Greene DN, Matsumura I (2006) Diversification and specialization of HIV protease function during in vitro evolution. Mol Biol Evol 23: 764-772.
16. Miyazaki K, Takenouchi M, Kondo H, Noro N, Suzuki M, et al. (2006) Thermal stabilization of Bacillus subtilis family-11 xylanase by directed evolution. J Biol Chem 281: 10236-10242.
17. Cedrone F, Menez A, Quemeneur E (2000) Tailoring new enzyme functions by rational redesign. Curr Opin Struct Biol 10: 405-410.
18. Lutz S (2010) Beyond directed evolution--semi-rational protein engineering and design. Curr Opin Biotechnol 21: 734-743.
19. Ni J, Takehara M, Watanabe H (2010) Identification of activity related amino acid mutations of a GH9 termite cellulase. Bioresour Technol 101: 6438-6443.
20. Edgar RC, Batzoglou S (2006) Multiple sequence alignment. Curr Opin Struct Biol 16: 368-373.
21. Notredame C (2007) Recent evolutions of multiple sequence alignment algorithms. PLoS Comput Biol 3: e123.
22. Pei J (2008) Multiple protein sequence alignment. Curr Opin Struct Biol 18: 382-386.
23. Feng DF, Doolittle RF (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J Mol Evol 25: 351-360.
24. Reese ML, Zeiner GM, Saeij JP, Boothroyd JC, Boyle JP (2011) Polymorphic family of injected pseudokinases is paramount in Toxoplasma virulence. Proc Natl Acad Sci U S A 108: 9625-9630.
25. Yao H, Kristensen DM, Mihalek I, Sowa ME, Shaw C, et al. (2003) An accurate, sensitive, and scalable method to identify functional sites in protein structures. J Mol Biol 326: 255-261.
26. Jacobs SA, Diem MD, Luo J, Teplyakov A, Obmolova G, et al. (2012) Design of novel FN3 domains with high stability by a consensus sequence approach. Protein Eng Des Sel 25: 107-117.
27. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673-4680.
28. Ehren J, Govindarajan S, Moron B, Minshull J, Khosla C (2008) Protein engineering of improved prolyl endopeptidases for celiac sprue therapy. Protein Eng Des Sel 21: 699-707.
29. Gumpena R, Kishor C, Ganji RJ, Jain N, Addlagatta A (2012) Glu121-Lys319 salt bridge between catalytic and N-terminal domains is pivotal for the activity and stability of Escherichia coli aminopeptidase N. Protein Sci 21: 727-736.
30. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, et al. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7: 539.
31. Notredame C, Higgins DG, Heringa J (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 302: 205-217.
32. Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30: 3059-3066.
33. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792-1797.
34. Raghava GP, Searle SM, Audley PC, Barber JD, Barton GJ (2003) OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy. Bmc Bioinformatics 4: 47.
35. Thompson JD, Linard B, Lecompte O, Poch O (2011) A comprehensive benchmark study of multiple sequence alignment

methods: current challenges and future perspectives. PLoS One 6: e18093.
36. Plyusnin I, Holm L (2012) Comprehensive comparison of graph based multiple protein sequence alignment strategies. Bmc Bioinformatics 13: 64.
37. Katoh K, Toh H (2008) Recent developments in the MAFFT multiple sequence alignment program. Brief Bioinform 9: 286-298.
38. Katoh K, Kuma K, Miyata T, Toh H (2005) Improvement in the accuracy of multiple sequence alignment program MAFFT. Genome informatics International Conference on Genome Informatics 16: 22-33.
39. Katoh K, Kuma K, Toh H, Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 33: 511-518.
40. Michel G, Pojasek K, Li Y, Sulea T, Linhardt RJ, et al. (2004) The structure of chondroitin B lyase complexed with glycosaminoglycan oligosaccharides unravels a calcium-dependent catalytic machinery. J Biol Chem 279: 32882-32896.
41. Maita N, Nyirenda J, Igura M, Kamishikiryo J, Kohda D (2010) Comparative structural biology of eubacterial and archaeal oligosaccharyltransferases. J Biol Chem 285: 4941-4950.
42. Ackerman SH, Gatti DL (2011) The contribution of coevolving residues to the stability of KDO8P synthase. PLoS One 6: e17459.
43. Lovell SC, Robertson DL (2010) An integrated view of molecular coevolution in protein-protein interactions. Mol Biol Evol 27: 2567-2575.
44. Hooper LV, Gordon JI (2001) Commensal host-bacterial relationships in the gut. Science 292: 1115-1118.
45. Chisholm ST, Coaker G, Day B, Staskawicz BJ (2006) Host-microbe interactions: shaping the evolution of the plant immune response. Cell 124: 803-814.
46. Moyle WR, Campbell RK, Myers RV, Bernard MP, Han Y, et al. (1994) Co-evolution of ligand-receptor pairs. Nature 368: 251-255.
47. Yeang CH, Haussler D (2007) Detecting coevolution in and among protein domains. PLoS Comput Biol 3: e211.
48. Wang ZO, Pollock DD (2007) Coevolutionary patterns in cytochrome c oxidase subunit I depend on structural and functional context. J Mol Evol 65: 485-495.
49. Wollenberg KR, Atchley WR (2000) Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. Proc Natl Acad Sci U S A 97: 3288-3291.
50. Saraf MC, Moore GL, Maranas CD (2003) Using multiple sequence correlation analysis to characterize functionally important protein regions. Protein Eng 16: 397-406.
51. Chakrabarti S, Panchenko AR (2009) Coevolution in defining the functional specificity. Proteins-Structure Function and Bioinformatics 75: 231-240.
52. Chaparro-Riggers JF, Polizzi KM, Bommarius AS (2007) Better library design: data-driven protein engineering. Biotechnol J 2: 180-191.
53. Chen Z, Meyer W, Rappert S, Sun J, Zeng AP (2011) Coevolutionary analysis enabled rational deregulation of allosteric enzyme inhibition in Corynebacterium glutamicum for lysine production. Appl Environ Microbiol 77: 4352-4360.
54. Zhang J, Rosenberg HF (2002) Complementary advantageous substitutions in the evolution of an antiviral RNase of higher primates. Proc Natl Acad Sci U S A 99: 5486-5491.
55. Perez-Jimenez R, Wiita AP, Rodriguez-Larrea D, Kosuri P, Gavira JA, et al. (2008) Force-clamp spectroscopy detects residue co-evolution in enzyme catalysis. J Biol Chem 283: 27121-27129.
56. Horner DS, Pirovano W, Pesole G (2008) Correlated substitution analysis and the prediction of amino acid structural contacts. Brief Bioinform 9: 46-56.
57. Gloor GB, Martin LC, Wahl LM, Dunn SD (2005) Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. Biochemistry 44: 7156-7165.
58. Atchley WR, Wollenberg KR, Fitch WM, Terhalle W, Dress AW (2000) Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. Mol Biol Evol 17: 164-178.
59. Brown CA, Brown KS (2010) Validation of coevolving residue algorithms via pipeline sensitivity analysis: ELSC and OMES and ZNMI, oh my! PLoS One 5: e10779.
60. Yip KY, Patel P, Kim PM, Engelman DM, McDermott D, et al. (2008) An integrated system for studying residue coevolution in proteins. Bioinformatics 24: 290-292.
61. Fodor AA, Aldrich RW (2004) Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. Proteins-Structure Function and Bioinformatics 56: 211-221.
62. Lockless SW, Ranganathan R (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. Science 286: 295-299.
63. Kass I, Horovitz A (2002) Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. Proteins-Structure Function and Bioinformatics 48: 611-617.
64. Gobel U, Sander C, Schneider R, Valencia A (1994) Correlated mutations and residue contacts in proteins. Proteins 18: 309-317.
65. Aurora R, Donlin MJ, Cannon NA, Tavis JE (2009) Genome-wide hepatitis C virus amino acid covariance networks can predict response to antiviral therapy in humans. J Clin Invest 119: 225-236.
66. Fuchs A, Martin-Galiano AJ, Kalman M, Fleishman S, Ben-Tal N, et al. (2007) Co-evolving residues in membrane proteins. Bioinformatics 23: 3312-3319.
67. del Sol A, Fujihashi H, Amoros D, Nussinov R (2006) Residues crucial for maintaining short paths in network communication mediate signaling in proteins. Mol Syst Biol 2: 2006 0019.
68. Dekker JP, Fodor A, Aldrich RW, Yellen G (2004) A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments. Bioinformatics 20: 1565-1572.
69. Halperin I, Wolfson H, Nussinov R (2006) Correlated mutations: advances and limitations. A study on fusion proteins and on the Cohesin-Dockerin families. Proteins 63: 832-845.
70. Viola RE (2001) The central enzymes of the aspartate family of amino acid biosynthesis. Acc Chem Res 34: 339-349.
71. Chen Z, Rappert S, Sun J, Zeng AP (2011) Integrating molecular dynamics and co-evolutionary analysis for reliable target prediction and deregulation of the allosteric inhibition of aspartokinase for amino acid production. J Biotechnol 154: 248-254.
72. Wei J, Zhou Y, Xu T, Lu B (2010) Rational design of catechol-2, 3-dioxygenase for improving the enzyme characteristics. Appl Biochem Biotechnol 162: 116-126.
73. Chen CY, Georgiev I, Anderson AC, Donald BR (2009) Computational structure-based redesign of enzyme activity. Proc Natl Acad Sci U S A 106: 3764-3769.
74. Fischer D (2006) Servers for protein structure prediction. Curr Opin Struct Biol 16: 178-182.
75. Zhang Y (2008) Progress and challenges in protein structure prediction. Curr Opin Struct Biol 18: 342-348.
76. Ginalski K (2006) Comparative modeling for protein structure prediction. Curr Opin Struct Biol 16: 172-177.
77. David R, Korenberg MJ, Hunter IW (2000) 3D-1D threading methods for protein fold recognition. Pharmacogenomics 1: 445-455.
78. Read RJ, Chavali G (2007) Assessment of CASP7 predictions in the high accuracy template-based modeling category. Proteins-Structure Function and Bioinformatics 69 Suppl 8: 27-37.

79. Jauch R, Yeo HC, Kolatkar PR, Clarke ND (2007) Assessment of CASP7 structure predictions for template free targets. Proteins-Structure Function and Bioinformatics 69 Suppl 8: 57-67.
80. Zhang Y (2009) Protein structure prediction: when is it useful? Curr Opin Struct Biol 19: 145-155.
81. Vitkup D, Melamud E, Moult J, Sander C (2001) Completeness in structural genomics. Nat Struct Biol 8: 559-566.
82. Mariani V, Kiefer F, Schmidt T, Haas J, Schwede T (2011) Assessment of template based protein structure predictions in CASP9. Proteins 79 Suppl 10: 37-58.
83. Hildebrand A, Remmert M, Biegert A, Soding J (2009) Fast and accurate automatic structure prediction with HHpred. Proteins 77 Suppl 9: 128-132.
84. Zhang Y (2008) I-TASSER server for protein 3D structure prediction. Bmc Bioinformatics 9: 40.
85. Xu D, Zhang J, Roy A, Zhang Y (2011) Automated protein structure modeling in CASP9 by I-TASSER pipeline combined with QUARK-based ab initio folding and FG-MD-based structure refinement. Proteins 79 Suppl 10: 147-160.
86. Rychlewski L, Fischer D (2005) LiveBench-8: the large-scale, continuous assessment of automated protein structure prediction. Protein Sci 14: 240-245.
87. Schwede T, Kopp J, Guex N, Peitsch MC (2003) SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Res 31: 3381-3385.
88. Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, et al. (2000) Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 29: 291-325.
89. Bordoli L, Schwede T (2012) Automated protein structure modeling with SWISS-MODEL Workspace and the Protein Model Portal. Methods Mol Biol 857: 107-136.
90. Lopatin AN, Nichols CG (2001) Inward rectifiers in the heart: an update on I(K1). J Mol Cell Cardiol 33: 625-638.
91. Ureche ON, Baltaev R, Ureche L, Strutz-Seebohm N, Lang F, et al. (2008) Novel insights into the structural basis of pH-sensitivity in inward rectifier K+ channels Kir2.3. Cell Physiol Biochem 21: 347-356.
92. Choi SH, Kim HS, Lee EY (2009) Comparative homology modeling-inspired protein engineering for improvement of catalytic activity of Mugil cephalus epoxide hydrolase. Biotechnol Lett 31: 1617-1624.
93. Qi X, Guo Q, Wei Y, Xu H, Huang R (2012) Enhancement of pH stability and activity of glycerol dehydratase from Klebsiella pneumoniae by rational design. Biotechnol Lett 34: 339-346.
94. Dehouck Y, Kwasigroch JM, Gilis D, Rooman M (2011) PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. Bmc Bioinformatics 12: 151.
95. Khan S, Vihinen M (2010) Performance of protein stability predictors. Hum Mutat 31: 675-684.
96. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, et al. (2005) The FoldX web server: an online force field. Nucleic Acids Res 33: W382-388.
97. Gilis D, Rooman M (2000) PoPMuSiC, an algorithm for predicting protein mutant stability changes: application to prion proteins. Protein Eng 13: 849-856.
98. Dehouck Y, Grosfils A, Folch B, Gilis D, Bogaerts P, et al. (2009) Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. Bioinformatics 25: 2537-2543.
99. Bava KA, Gromiha MM, Uedaira H, Kitajima K, Sarai A (2004) ProTherm, version 4.0: thermodynamic database for proteins and mutants. Nucleic Acids Res 32: D120-121.
100. Smith BA, Hecht MH (2011) Novel proteins: from fold to function. Curr Opin Chem Biol 15: 421-426.
101. Kaplan J, DeGrado WF (2004) De novo design of catalytic proteins. Proc Natl Acad Sci U S A 101: 11566-11570.
102. Reina J, Lacroix E, Hobson SD, Fernandez-Ballester G, Rybin V, et al. (2002) Computer-aided design of a PDZ domain to recognize new target sequences. Nat Struct Biol 9: 621-627.
103. Looger LL, Dwyer MA, Smith JJ, Hellinga HW (2003) Computational design of receptor and sensor proteins with novel functions. Nature 423: 185-190.
104. Garcia-Viloca M, Gao J, Karplus M, Truhlar DG (2004) How enzymes work: analysis by modern rate theory and computer simulations. Science 303: 186-195.
105. Tantillo DJ, Chen J, Houk KN (1998) Theozymes and compuzymes: theoretical models for biological catalysis. Curr Opin Chem Biol 2: 743-750.
106. Richter F, Leaver-Fay A, Khare SD, Bjelic S, Baker D (2011) De novo enzyme design using Rosetta3. PLoS One 6: e19230.
107. Dechancie J, Clemente FR, Smith AJ, Gunaydin H, Zhao YL, et al. (2007) How similar are enzyme active site geometries derived from quantum mechanical theozymes to crystal structures of enzyme-inhibitor complexes? Implications for enzyme design. Protein Sci 16: 1851-1866.
108. Zanghellini A, Jiang L, Wollacott AM, Cheng G, Meiler J, et al. (2006) New algorithms and an in silico benchmark for computational enzyme design. Protein Sci 15: 2785-2794.
109. Kuhlman B, Baker D (2000) Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci U S A 97: 10383-10388.
110. Kiss G, Rothlisberger D, Baker D, Houk KN (2010) Evaluation and ranking of enzyme designs. Protein Sci 19: 1760-1773.
111. Altamirano MM, Blackburn JM, Aguayo C, Fersht AR (2000) Directed evolution of new catalytic activity using the alpha/beta-barrel scaffold. Nature 403: 617-622.
112. Alexandrova AN, Rothlisberger D, Baker D, Jorgensen WL (2008) Catalytic mechanism and performance of computationally designed enzymes for Kemp elimination. J Am Chem Soc 130: 15907-15915.
113. Khersonsky O, Rothlisberger D, Dym O, Albeck S, Jackson CJ, et al. (2010) Evolutionary optimization of computationally designed enzymes: Kemp eliminases of the KE07 series. J Mol Biol 396: 1025-1042.
114. Khersonsky O, Rothlisberger D, Wollacott AM, Murphy P, Dym O, et al. (2011) Optimization of the in-silico-designed kemp eliminase KE70 by computational design and directed evolution. J Mol Biol 407: 391-412.
115. Bolon DN, Mayo SL (2001) Enzyme-like proteins by computational design. Proc Natl Acad Sci U S A 98: 14274-14279.
116. Nat Chem BiolPLoS MedFaiella M, Andreozzi C, de Rosales RT, Pavone V, Maglio O, et al. (2009) An artificial di-iron oxo-protein with phenol oxidase activity. Nat Chem Biol 5: 882-884.
117. Yeung N, Lin YW, Gao YG, Zhao X, Russell BS, et al. (2009) Rational design of a structural and functional nitric oxide reductase. Nature 462: 1079-1082.
118. Dahiyat BI, Mayo SL (1997) De novo protein design: Fully automated sequence selection. Science 278: 82-87.
119. Bornscheuer UT, Huisman GW, Kazlauskas RJ, Lutz S, Moore JC, et al. (2012) Engineering the third wave of biocatalysis. Nature 485: 185-194.
120. Ioannidis JP (2005) Why most published research findings are false. PLoS Med 2: e124.
121. Lippow SM, Tidor B (2007) Progress in computational protein design. Curr Opin Biotechnol 18: 305-311.

122. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, et al. (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23: 2947-2948.
123. Poirot O, O'Toole E, Notredame C (2003) Tcoffee@igs: A web server for computing, evaluating and combining multiple sequence alignments. Nucleic Acids Res 31: 3503-3506.
124. Arnold K, Bordoli L, Kopp J, Schwede T (2006) The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics 22: 195-201.
125. Kiefer F, Arnold K, Kunzli M, Bordoli L, Schwede T (2009) The SWISS-MODEL Repository and associated resources. Nucleic Acids Res 37: D387-392.
126. Roy A, Kucukural A, Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 5: 725-738.
127. Guerois R, Nielsen JE, Serrano L (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J Mol Biol 320: 369-387.
128. Kwasigroch JM, Gilis D, Dehouck Y, Rooman M (2002) PoPMuSiC, rationally designing point mutations in protein structures. Bioinformatics 18: 1701-1702.
129. Capriotti E, Fariselli P, Rossi I, Casadio R (2008) A three-state prediction of single point mutations on protein stability changes. Bmc Bioinformatics 9 Suppl 2: S6.
130. Zhou H, Zhou Y (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 11: 2714-2726.