COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

Volume No: 2, Issue: 3, September 2012, e201209016, http://dx.doi.org/10.5936/csbj.201209016

CSBJ

Computational tools for rational protein engineering of aldolases

Michael Widmann [a], Jürgen Pleiss [a], Anne K. Samland [b,*]

Abstract: In this mini-review we describe the different strategies for rational protein engineering and summarize the computational tools available. Computational tools can either be used to design focused libraries, to predict sequence-function relationships or for structure-based molecular modelling. This also includes *de novo* design of enzymes. Examples for protein engineering of aldolases and transaldolases are given in the second part of the mini-review.

Mini Review Article

1. Introduction

Asymmetric aldol additions are a corner stone of preparative organic chemistry. Concomitant with the formation of a C-C bond between a nucleophile (donor) and an electrophile (acceptor) one or two new stereocenters are created. This type of reaction can also be carried out by enzymes, such as aldolases and transaldolases. Those enzymes, in most cases, strictly control the stereo configuration at the newly formed stereocenter(s). Aldolases are applied in biocatalysis for the synthesis of amino acid and carbohydrate derivatives. For more details about aldolases and their biocatalytic application see recent reviews [1-4].

Mechanistically, class I and class II aldolases are distinguished. Class I aldolases form a Schiff base intermediate between a conserved Lys in the active site and the carbonyl carbon atom of the donor substrate, i.e. usually a ketone. By proton abstraction an enamine intermediate is formed which attacks the carbonyl carbon atom of the acceptor aldehyde. Class I aldolases do not require any cofactor and they exhibit a typical $(\beta/\alpha)_8$-barrel fold. Class II aldolases depend on a divalent cation which acts as a Lewis acid. The metal ion helps to deprotonate the donor substrate and stabilises the enolate formed. Therefore, these aldolases can be inhibited by EDTA. According to their structure and sequence class I and class II aldolases do not show any significant homology. Apparently, they evolved separately.

Aldolases usually accept a wide range of acceptor substrates which allows a broad range of synthetic applications. On the other hand, they are in general very specific for their donor substrate. Hence, they are classified as (i) dihydroxyacetone phosphate (DHAP) dependent aldolases, (ii) dihydroxyacetone (DHA) dependent aldolases, (iii) pyruvate/2-oxobutyrate dependent aldolases, (iv) acetaldehyde dependent aldolases and (v) glycine/alanine dependent aldolases [1]. Glycine/alanine dependent aldolases are neither class I nor class II aldolase but require pyridoxal phosphate (PLP) as cofactor. Structurally, they belong to the fold type I family of PLP dependent enzymes.

Transaldolases (Tal) transfer a DHA moiety from a ketose donor to an aldehyde acceptor. A new C-C bond is formed with 3*S*,4*R* stereo configuration. Mechanistically (Schiff base intermediate) and structurally ($(\beta/\alpha)_8$-barrel fold), Tals show similarity to class I aldolases. However, compared to DHAP dependent class I aldolases the conserved Lys residue moved to a different β-strand suggesting a circular permutation of the protein sequence [5]. Tals are almost ubiquitous enzymes and according to their sequence similarity they were divided into five subfamilies. The wild type enzyme did not find much application in biocatalysis. For more details on the Tal enzyme family see recent publications [6, 7].

Using computational tools protein engineering within this enzyme family was directed towards the following aims: (i) the discovery of new enzymes, (ii) the differentiation between enzyme families or subfamilies, (iii) the engineering of enzymes for new applications and (iv) the design of novel aldolases. In this mini-review we will first describe the different strategies for protein engineering and summarize the computational tools available. In the second part, we will give examples from the enzyme family of aldolases and transaldolases.

2. Computational tools for protein engineering

Isolated enzymes have been successfully applied for bioconversions provided the enzyme is stable, soluble, and easy to produce. However, in most cases the commercially available enzymes are not optimal for the desired chemical process. Therefore, *in silico*, *in vitro*, and *in vivo* strategies have been developed to screen for appropriate enzymes from the natural pool [8]. However, natural enzymes rarely have the combined properties necessary for industrial chemical production such as high activity, high selectivity, broad substrate specificity towards non-natural substrates, no inhibition by substrate or product, and a high stability in organic solvents and at high substrate or product concentrations [9]. Therefore, protein engineering has been successfully applied to design enzymes with new substrate spectra and new functions as catalysts for unnatural substrates, and to fine-tune bottleneck enzymes in metabolic engineering [10]. Three major computational strategies are currently applied to support protein engineering: directed evolution, methods to predict sequence-function relationships, and structure-based molecular modelling methods.

[a]*Institute of Technical Biochemistry, University of Stuttgart, Allmandring 31, 70569 Stuttgart, Germany*

[b]*Institute of Microbiology, University of Stuttgart, Allmandring 31, 70569 Stuttgart, Germany*

* Corresponding author. Tel.: +49 71168565491; Fax: +49 71168565725

E-mail address: anne.samland@imb.uni-stuttgart.de (Anne K. Samland)

2.1 Design of focused libraries for directed evolution

Directed evolution has proven to be an effective method to improve the properties of enzymes (for aldolases see review [11]). The unguided use of random mutagenesis methods, however, results in protein libraries with millions of members which still only sample a small fraction of the vast sequence space possible [12]. Recently, several computational approaches have been suggested to improve the efficiency of the directed evolution by enriching the library and reducing the library size substantially, taking into account further information. An enrichment of the library may be achieved by considering structure information on residues that are involved in substrate binding. This approach has guided the design of highly focused libraries and resulted in mutants with increased selectivity [13-15] or shifted substrate specificity [16-19]. The size of the library can be reduced by limiting the possible amino acid alphabet, i.e. not all 20 amino acids but a subset is used instead, depending on the desired interactions [20]. To estimate the screening effort necessary the CASTER tool was developed by the Reetz group. A comprehensive statistical analysis of a large number of favourable and less favourable mutants identified hot spot regions that are beneficial to enzyme activity and stability [21-23]. Most of these methods to search for promising mutation sites require expert knowledge in bioinformatics which may not be present in experimentally oriented research groups. Therefore, online tools that require little to none bioinformatics knowledge have become popular. Meta-tools such as the HotSpot Wizard [24] offer a complete workflow to assess promising mutation sites by combining a variety of methods such as Catalytic Site Atlas [25], CASTp [26], CAVER [27], BLAST [28], MUSCLE [29], as well as sequence and structure databases such as UniProt [30], NCBI GenBank [31], and PDB [32].

2.2 Prediction of sequence-function relationships

The second strategy takes advantage of the rapidly growing amount of available protein sequences, structures, functional and biochemical data. Systematic analyses are based on large number of protein sequences and complete protein families to yield insights into catalytic mechanisms and evolutionary pathways [33]. By comparing the sequences of homologous proteins, consensus or ancestor sequences were constructed. Back-to-the-consensus mutations were shown to increase stability [34-36] or improve expression [37]. Recently, ancestral mutations have been integrated with directed evolution to generate a stabilized starting point of highly diverse and evolvable gene libraries [38]. Alternatively, multi-sequence alignments were analyzed to identify correlated mutations, to identify structurally or functionally relevant residues [39, 40], and to predict mutants with improved substrate specificity, catalytic activity, or protein stability [41]. Sequence-based methods were also applied to predict aggregation-prone regions [42] and to design mutants with decreased aggregation rates [43]. Multiple sequence alignments assisted by structural information were also used to identify subfamily specific positions in aldolases [44-46].

While the amount of information on sequence, structure, and biochemical information is steadily increasing, it is generally not available to a systematic analysis. Therefore, databases have been developed that provide access to enzymatic information such as BRENDA [47] or to integrate information on enzyme families such as DWARF [48] and 3DM [49]. BRENDA (BRaunschweig ENzyme DAtabase) offers a comprehensive collection of biochemical data on a broad range of enzyme families, which are grouped according to their EC numbers, providing information about reaction type, products, and substrates, organisms of origin, and an overview of available publications. The DWARF system (Data Warehouse system for Analyzing pRotein Families) integrates sequence, structure, and annotation information of large protein families including lipases [50], triterpene cyclases [51], thiamine-diphosphate dependent enzymes [52], and lactamases [53]. The 3DM system [54] is based on the creation of structure-based multiple sequence alignments. A common numbering scheme for structurally equivalent amino acids allows for the automated creation of homology models, the analysis of correlated or conserved residues and the prediction of functionally relevant residues [41, 55]. As of the time of this review, no database with a focus on aldolases has been published.

2.3 Structure-based molecular modelling

The third strategy starts from information on protein structure and seeks to improve stability, activity, specificity, or selectivity by molecular modelling. While for a growing number of proteins, experimentally determined structure information become available by the Protein Data Bank [32], only for a small fraction of all proteins with known sequence the structure is also known. However, if sequence similarity is sufficiently high the structure of a protein can be modeled based on a sequence comparison to a protein with experimentally determined structure. Sequence identities as low as 25% are usually enough to predict reliable structure models, in some cases even sequences with lower sequence identities are suitable for homology modeling [56]. Homology modeling programs such as Swiss-Model [57], Modeller [58] or Rosetta [59] are based on the observation that during evolution structure has been more conserved than sequence. Thus, proteins with similar sequence have a similar structure. Using these methods, structure models can be derived for the majority of soluble proteins as demonstrated by the biannual Critical Assessment of Protein Structure Prediction [60].

Many strategies for protein stabilization have been proposed: optimization of the distribution of surface charge–charge interactions [61, 62], improvement of core packing [63] and of the protein surface [64], and rigidification by introduction of prolines, exchange of glycines, introduction of disulfide bridges [65] or mutagenesis at positions with high B-factor [66]. However, it is still challenging to reliably predict mutations that stabilize the enzyme without affecting its activity or selectivity, which are a direct consequence of the molecular recognition of the substrate by the enzyme. For a change in stereoselectivity the side chains in vicinity of the stereocentre can be determined from structural data. These residues can then be split into sectors containing two to three residues which are randomized simultaneously [67, 68]. To improve activity and selectivity, modelling of the enzyme-substrate complex by molecular docking methods has been used to study the molecular basis of specificity and selectivity, and to predict mutations in the enzyme or modifications of the substrate structure that mediate specificity or selectivity [69-71]. It is recognized that shape and physico-chemical properties of the active site and the substrate binding site are the major driving forces to provide the specific interactions between enzyme and the transition state of the substrate that lead to catalysis. Moreover, there is increasing evidence that flexibility of the enzyme-substrate complex is crucial to recognition, because minor structural adjustments can have a big impact on the docking score [51]. Docking has been extensively used to predict substrate specificity and to identify positions that mediate substrate binding. Amino acids that clash with the desired substrate upon docking were exchanged, leading to an increase of catalytic activity of the enzyme variant toward this substrate [72-74]. Catalytic activity is mediated by only a small number of amino acids, metals, or cofactors located in the vicinity of the active site. However, substrate specificity and selectivity of an enzyme might be determined by factors beyond the geometric shape of the active site, such as long-