PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Correlated mutation analyses on super-family alignments reveal functionally important residues

Remko K. P. Kuipers,[1] Henk-Jan Joosten,[2] Eugene Verwiel,[2] Sjoerd Paans,[2] Jasper Akerboom,[2] John van der Oost,[2] Nicole G. H. Leferink,[3] Willem J. H. van Berkel,[3] Gert Vriend,[1] and Peter J. Schaap[2]*

[1] Centre of Molecular and Biomolecular Informatics, Radboud University, Nijmegen, The Netherlands
[2] Laboratory of Microbiology, Wageningen University, Wageningen, The Netherlands
[3] Laboratory of Biochemistry, Wageningen University, Wageningen, The Netherlands

ABSTRACT

Correlated mutation analyses (CMA) on multiple sequence alignments are widely used for the prediction of the function of amino acids. The accuracy of CMA-based predictions is mainly determined by the number of sequences, by their evolutionary distances, and by the quality of the alignments. These criteria are best met in structure-based sequence alignments of large super-families. So far, CMA-techniques have mainly been employed to study the receptor interactions. The present work shows how a novel CMA tool, called Comulator, can be used to determine networks of functionally related residues in enzymes. These analyses provide leads for protein engineering studies that are directed towards modification of enzyme specificity or activity. As proof of concept, Comulator has been applied to four enzyme super-families: the isocitrate lyase/phoshoenol-pyruvate mutase super-family, the hexokinase super-family, the RmlC-like cupin super-family, and the FAD-linked oxidases super-family. In each of those cases networks of functionally related residue positions were discovered that upon mutation influenced enzyme specificity and/or activity as predicted. We conclude that CMA is a powerful tool for redesigning enzyme activity and selectivity.

Proteins 2009; 76:608–616.

Key words: comulator; 3DM; protein engineering; rational design; hexo-kinases; isocitrate-lyase/phosphoenolpyruvate lyases; cupins; FAD-oxidases.

INTRODUCTION

Proteins evolve within a framework of functional constraints that limit substitutions at individual positions in the sequence. The results of these constraints can be detected in large multiple sequence alignments (MSAs) as evolutionary fingerprints. Coevolution of the amino acids at two distinct alignment positions, for example, is a result of functional constraints that force compensating mutations for specific residue changes. This coevolution of residue positions can be detected by correlated mutation analyses (CMA) algorithms. Although the concept of correlated mutations is rather straightforward, their unambiguous detection proved more difficult. Therefore, several algorithms have been developed that are able to screen alignments for correlated mutations.[1,2] These methods are mostly used for the prediction of contacts between residues. Contact predictions can reveal intermolecular protein–protein interactions,[3–5] or intramolecular interactions that in turn can be used for protein structure predictions.[6] In 1993 we introduced the idea that CMA is better suited for the detection of functionally related residues.[7] Later, using the GPCR protein super-family, we showed that residue positions with common function indeed tend to stay conserved, and when they do change they do so simultaneously.[8] Nevertheless, the number of articles that describe the utilization of CMA for detection of functionally related residues is still very limited. There are some examples where CMA was used to identify ligand–receptor interactions sites[8–10] and recently two articles appeared in which CMA was successfully used to detect residue positions important in multidrug resistance of the HIV-1 protease.[10,11]

Unambiguous detection of functionally related residues by CMA requires a reliable, large super-family alignment. We have recently designed the 3DM software (manuscript in preparation) that can be used to rapidly produce structure-based super-family MSAs. The Comulator software is a novel extension of this 3DM software suite, and was specifically designed for the analysis of very large MSAs. Here it was used in

Grant sponsor: Netherlands Bioinformatics Centre (NBIC).
*Correspondence to: Peter J. Schaap, Laboratory of Microbiology, Dreijenplein 10, 6703 HB Wageningen, The Netherlands. E-mail: peter.schaap@wur.nl.
Received 9 October 2008; Revised 10 December 2008; Accepted 19 December 2008
Published online 20 January 2009 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.22374

protein engineering experiments to analyze the alignments of four super-families: (1) the isocitrate lyase/phosphoenol-pyruvate mutase (ICL/PEPM) super-family, in which we could selectively remove the specificity for one of its substrates, (2) the hexokinase (HK) super-family, in which we could successfully predict the allowed subset of allowed residues in a saturation mutagenesis experiment, (3) the FAD-linked oxidases (FAD-O) super-family in which we could predict compensating mutations for loss-of-function mutants, and (4) the RmlC-like cupin (cupin) super-family (nomenclature according to the SCOP database[12]), in which we could improve the activity by designing a double mutant. The former two super-family MSAs were used to detect residues involved in substrate specificity, the latter two to predict compensating mutations for mutations that either decreased protein activity or protein stability. The Comulator CMA results agree well with experiments in all four cases; the latter two sets of experiments were produced by us and are published here for the first time.

The Comulator is available at http://3dmcsis.systemsbiology.nl/comulator/; this most likely is the first CMA software that is freely available via the internet. Structure and sequence alignments, CMA results, mutations mined from literature, and so forth, for the here investigated four super-families is available at http://3dmcsis.systemsbiology.nl/.

MATERIALS AND METHODS

Super-family sequence alignment

Protein structures belonging to four super-families were collected using the SCOP database combined with BLAST[13] searches in the PDB database. Super-family sequences were collected by BLAST searches in the NCBI database using the sequences of the super-family structures as query sequences. The super-family alignments were generated using 3DM. This software is described elsewhere (manuscript in preparation) and is only briefly described here. 3DM superposes the structures of proteins belonging to a super-family and so generates a structure-based sequence alignment that contains all sequence positions that are structurally conserved throughout the super-family. These so-called core positions are numbered sequentially, and these numbers (called 3DM-numbers) are used throughout this study. For each structure a sequence profile is built by iteratively aligning protein sequences that are at least 30% identical to protein sequence of the structure. Next, these sequences are aligned against the profile they are most similar to. In the final step the superimposed structures are used as guidance to merge individual alignments into one large super-family alignment.

CM algorithm

The Comulator algorithm is derived from a method described for the detection of allosteric interactions in the nuclear receptor super-family.[14] The underlying method is known as the statistical coupling analysis method.[15,16] Equation 1 shows the similar method that was implemented in the Comulator.

$$CM(x,y) = \sum_{a=1}^{20}\sum_{b=1}^{20} |F(x,y,a,b)|\; N(x=a \wedge y=b)/400N$$

with,

$$F(x,y,a,b) = \frac{N(y=b)}{N} - \frac{N(x=a \wedge y=b)}{N(x=a)} \qquad (1)$$

In which x and y run over the residue positions in the MSA; $CM_{(x,y)}$ is the correlation score between the residue positions x and y; a and b run over the 20 amino acid types; N is the number of sequences in the MSA; $N_{(y=b)}$ is the frequency of residue type b at position y; $N_{(x=a)}$ is the frequency of residue type a at position x; and $N_{(x=a\wedge y=b)}$ is the frequency of residue type b at position y in sequences where type a is observed at position x; $|F_{(x,y,a,b)}|$ is the absolute value of $F_{(x,y,a,b)}$. F will be negative if the amino acid b at position y is relatively more abundant in the subset of sequences that has type a at position x $[N_{(x=a\wedge y=b)}/N_{(x=a)}]$ than in the full alignment at position y $[N_{(y=b)}/N]$. If the amino acid pair is observed less often in the subset than in the whole alignment the score F will be positive. Due to the summation of absolute values, pairs of residue types contribute more to the score $CM_{(x,y)}$ when their frequency deviates more from the average. In other words, a high F score is obtained when the residues at positions x and y tend to mutate in tandem. Obviously, a pair of fully conserved positions gets a score of zero. Comulator calculates CM scores for all possible alignment position pairs. The resulting scores are visualized in heat-maps that are incorporated in interactive HTML pages (see Fig. 1) in which all squares are hyperlinked to the underlying raw data, including the alignment.

Comulator website

The WWW based version of Comulator accepts as input aligned sequences in Fasta or ClustalW format. The input alignments are visualized similarly as for 3DM-derived alignments including an alignment positions numbering scheme. The same numbering scheme is applied to the CM heat-map. If a sequence file contains a Swiss-Prot ID then the results are also automatically linked to the corresponding Swiss-Prot data file.

Mutagenesis, over-expression and purification of phospho-glucose isomerase from *Pyrococcus furiosus*

The cloning of *pgiA* is described by Verhees *et al.*[17] Mutants were generated with the QuikChange Site-Directed

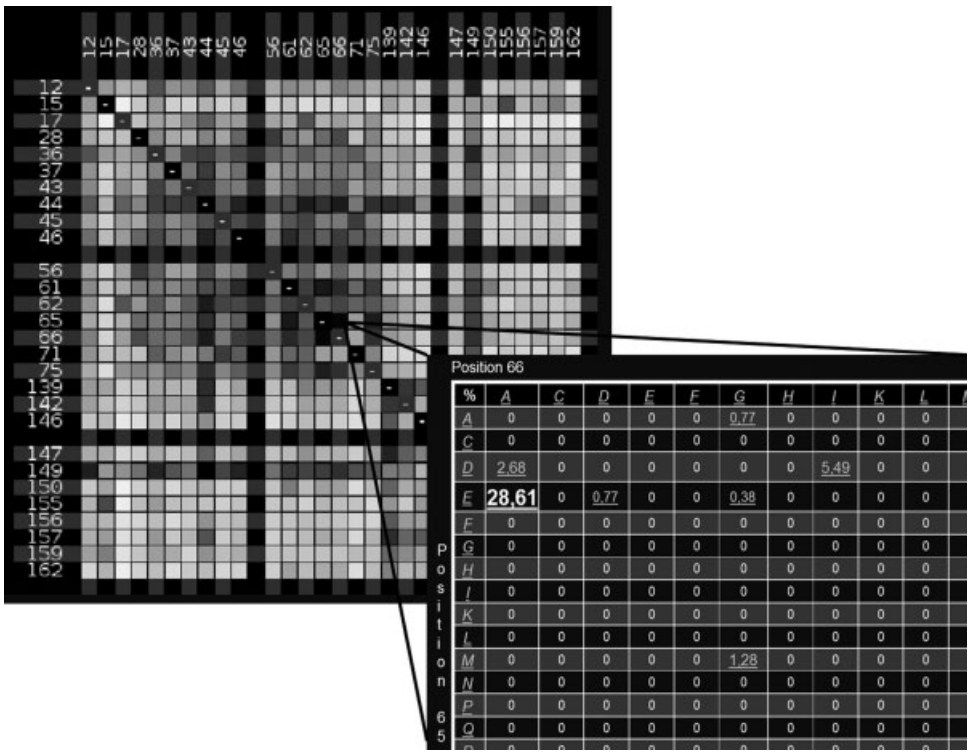

Figure 1

Heat-map of the ICL/PEPM super-family. Only residue positions are shown that have at least one correlation score above a cut-off (CMA score >0.8). Squares are colored from green (low correlation) to red (high correlation). The inset shows an example of hyperlinked information. In the insert the top-left corner of the residue pair frequency table is shown for the position pair.

Mutagenesis Kit (Stratagene) following the manufacturer's instructions with the following adaptations: 25 PCR cycles were applied, and the PCR mixture was incubated with *Dpn*I for 4–8 h at 37°C. Mutants and primers used for mutagenesis are listed in Table I. Mutants were verified by sequencing (Baseclear, Leiden, The Netherlands).

Escherichia coli strain BL21(DE3) containing the tRNA accessory plasmid pRIL (Stratagene) carrying the concerning plasmid was routinely grown in 1 L Luria Bertani medium (LB-medium) with kanamycin and chloramphenicol at 37°C until an OD_{600} of 0.5 was obtained. Isopropyl-β-D-thiogalactopyranoside (IPTG) was added to a final concentration of 0.1 m*M* and the culture was further incubated for 8 h under the same conditions. Cells were harvested by centrifugation (3800*g* at 4°C for 20 min), resuspended in 10 mL lysis buffer (20 m*M* Tris HCl, pH 8.0) and sonicated for 5 min at 4°C. The cell extract was clarified by centrifugation (37,000*g* at 4°C for 20 min). *Escherichia coli* proteins were denatured by incubating the cell extract at 70°C for 30 min, and pelleted by centrifugation (37,000*g* at 4°C for 20 min). PGI was purified to homogeneity using FPLC: the supernatant was loaded onto a Q-Sepharose Fast Flow column (GE Healthcare)

Table I

Primers Used for the Mutagenesis Studies of pgiA

Mutation		
3DM #	PfPGI #	QuikChange primers
P27A	P132A	FW: (5′-GTAGTTTATGTTCCCGCCTATTGGG CTCATAGG-3′) RV: (5′-CCTATGAGCCCAATAGGCGGGAAC ATAAACTAC-3′)
Y28G	Y133G	FW: (5′-GTAGTTTATGTTCCCCCCGGTTGGG CTCATAGGACGG-3′) RV: (5′-CCGTCCTATGAGCCCAACCGGGGGG AACATAAACTAC-3′)
P27A/Y28G	P132A/Y133G	FW: (5′-GTAGTTTATGTTCCCGCCGGTTGGG CTCATAGGACGG-3′) RV: (5′-CCGTCCTATGAGCCCAACCGGCGGG AACATAAACTAC-3′)
P27E/Y28G	P132E/Y133G	FW: (5′-GTAGTTTATGTTCCCGAAGGTTGGG CTCATAGGACGG-3′) RV: (5′-CCGTCCTATGAGCCCAACCTTCGGG AACATAAACTAC-5′)
P27R/Y28G	P132R/Y133G	FW: (5′-GTAGTTTATGTTCCCCGCGGTTGGG CTCATAGGACGG-3′) RV: (5′-CCGTCCTATGAGCCCAACCGCGGGG AACATAAACTAC-5′)

Both the 3DM alignment position numbering (first column) and the number of the corresponding position in the ORF of phoshoglucose isomerase from *P. furiosus* (second column) are indicated.

Table II
Primers Used for the Mutagenesis Studies of AtGALDH

Mutation		
3DM#	AtGALDH#	QuikChange primers
36	L56H	FW: (5′-CCCGTTGGATCGGGTCACTCGCCTAATGGGATTG-3′) RV: (5′-CAATCCCATTAGGCGAGTGACCCGATCCAACGGG-3′)
78	A113G	FW: (5′-CTCTTCAGAACTTTGGCTCCATTAGAGAGCAG-3′) RV: (5′-CTGCTCTCTAATGGAGCCAAAGTTCTGAAGAG-3′)
91	V126G	FW: (5′-GGTGGTATTATTCAGGGTGGGGCACATGGGAC-3′) RV: (5′-GTCCCATGTGCCCCACCCTGAATAATACCACC-3′)

Both the 3DM alignment position numbering (first column) and the number of the corresponding position in the ORF of L-galactono-1,4-lactone dehydrogenase from *Arabidopsis thaliana* (second column) are indicated.

pre-equilibrated with 20 m*M* Tris HCl (pH 8.0). Proteins were eluted by a linear gradient of 0.0–1.0*M* NaCl in 20 m*M* Tris HCl (pH 8.0). Fractions containing PGI were pooled, concentrated, and loaded on a Superdex 200 GL column running in 20 m*M* Tris HCl containing 125 m*M* NaCl. Enzyme activity of the PGI mutants with fructose 6-phosphate was determined at 50°C as described previously[17] with the following adaptations: 20 m*M* Tris HCl pH 7.0 was used, and the protein samples were preincubated with 50 m*M* EDTA at 50°C for 20 min to ensure complete metal depletion.[18] Activity was measured with $MnCl_2$ in excess over EDTA to ensure enzyme saturation.

Expression, purification, and mutagenesis of L-galactono-1,4-lactone dehydrogenase from *Arabidopsis thaliana*

The cDNA encoding mature L-galactono-1,4-lactone dehydrogenase (GALDH) from *A. thaliana* has been cloned previously to yield pET-GALDH-His_6[19]. The GALDH mutants used in this study were constructed using pET-GALDH-His_6 as template with the QuikChange method (Stratagene) using the primers listed in Table II. Successful mutagenesis was confirmed by automated sequencing.

For enzyme production *E. coli* BL21(DE3) cells, harboring a pET-GALDH plasmid, were grown in 1 L LB-medium supplemented with 100 μg/mL ampicillin at 37°C until an OD_{600} of 0.7 was reached. Expression was induced by addition of 0.4 m*M* IPTG and the incubation was continued for 16 h at 37°C. The cells were harvested by centrifugation, resuspended in 5 mL lysis buffer (50 m*M* sodium phosphate, 300 m*M* NaCl, pH 7.4) and passed twice through a precooled French Press (SLM Aminco) at 10,000 PSI. The resulting homogenate was centrifuged at 25,000*g* for 30 min at 4°C to remove cell debris and the supernatant was loaded onto a HisGraviTrap column (GE Healthcare), equilibrated with 50 m*M* sodium phosphate, 300 m*M* NaCl, 45 m*M* imidazole, pH 7.4. Proteins were eluted with 50 m*M* sodium phosphate, 300 m*M* NaCl, 300 m*M* imidazole, pH 7.4 and saturated with FAD. Excess FAD and salt were removed by Biogel P-6DG size exclusion chromatography (BioRad) in 20 m*M* sodium phosphate, 0.1 m*M* DTT, pH 7.4. The amount of protein-bound FAD was determined from the ratio in absorbance at 280 and 450 nm (F-factor).

GALDH activity was routinely assayed by following the reduction of cytochrome *c* at 550 nm using a molar difference absorption coefficient ($\Delta\varepsilon_{550}$) of 21 mM^{-1} cm^{-1} for reduced minus oxidized cytochrome *c* as described,[19] with the modification that 1 μ*M* FAD was included in the assay mixture. The thermal stability of GALDH was determined as reported earlier.[19]

RESULTS AND DISCUSSION

Isocitrate lyase-like/Phosphoenolpyruvate mutase super-family

The isocitrate lyase/PEP mutase super-family alignment contains enzymes that cover three of the six main enzyme families (EC numbers) that all break a carbon–carbon in an oxalate-containing compound. All enzymes in this super-family share an α-β-barrel fold. The structure-based MSA contains 375 unique sequences. A network of nine residue positions with high CM scores was detected. These nine residue positions are found mainly but not exclusively surrounding the active site cleft. The detailed function of many of these residues is not yet known. However, these CM scores led to the discovery of a serine that is very specific for the oxaloacetate hydrolase (OAH) subfamily.[20] Mutating this serine to alanine, threonine, or proline (the most prevalent residues in other sequences in the alignment) did not significantly decrease the activity (k_{cat}), but had drastic effects on the affinity of OAH for its substrate oxaloacetate. This serine was used as a marker for family members with OAH activity, distinguishing OAH's from closely related paralogs that cannot convert oxaloacetate.[20] To show that this residue is indeed crucial for oxaloacetate recognition rather than for enzyme activity, we mutated this serine in the homologs petal death protein. This protein can convert a broad range of substrates including oxaloacetate. Indeed, when we mutated this serine we observed a 100-fold decrease in affinity for oxaloacetate, while the affinities for other substrates remained virtually unaffected.[20] Although OAH is very specific for oxaloacetate, it does have the potential to convert 2*R*,3*S*-2,3-dimethylmalate (DMM), albeit with poor efficiency. The S157A and S157P mutations in OAH actually improved the affinity of OAH for this substrate.[20] Recently we have isolated and characterized dimethyl-malate-lyase (DMML). DMML is closely related to OAH and has a proline instead of a serine at position 157. Mutating this proline

Figure 2

Tube presentation of a HK complexed with glucose (PDB accession code: 1BDG). Conserved residues (>90%) are shown in yellow and the network of highly correlating residues is shown in purple. The numbers assigned to the correlating positions are according to the 3DM numbering scheme. The glucose substrate is shown as balls model.

to a serine shows the same behavior with respect to k_{cat} and affinity.[21]

Hexokinase super-family

The sequences of the HK super-family can be divided into two main groups: the HKs that can phosphorylate a wide range of hexo-sugars, and glucokinases that specifically phosphorylate glucose. The Comulator detected a network of six highly correlated residue positions (3DM-numbers 46, 47, 52, 67, 82, and 88) that surround the active site (see Fig. 2). In humans/mammals glucokinases are HKs that often function in the liver. They are highly specific for glucose and work at high glucose concentrations. The "other" HKs tend to have a broad range of substrates (e.g., alose, mannose, or glucosamine) and are observed in a wide variety of cell-types. In contrast to glucokinases, they function well at low substrate concentrations. The Comulator found a network of highly correlated positions that contact the substrate. 3DM-supported visual inspection of the 709 HK sequences revealed that these six residue positions were conserved among the glucokinases. They were also conserved, but different among all "other" HKs. The observed correlations almost perfectly separate the two main groups in the super-family. The fingerprints for these two groups are 46[A]47[C,G,N]52[N]67[F]82[G]88[G] for the glucokinases, and 46[S]47[F,Y] 52[K]67[T]82[I]88[N] for the HKs. This clean separation suggests that these residue positions play a key role in determining the substrate specificities. Four of these positions (46, 47, 67, and 88) have been subject of a saturation mutagenesis experiment on the glucokinase of *E. coli* in a study by Miller.[22] *In vivo* selection in a glucokinase-deficient strain was used to find allowed substitutions at these positions. We analyzed all sequences in the MSA that have at least four of these six fingerprint residues in agreement with the glucokinase consensus [A,(CGN),N,F,G,G]. Most of these sequences possess consensus residues at all six positions, but about 10% differed at one or two positions. Despite the relatively low percentage of mutations, the results are still statistically meaningful because of the massive number of sequences used in this study. Table III lists the residue types that were observed at these six fingerprint positions. The saturation mutagenesis experiments that were performed for four of these six residues are also shown. In 65 of 76 mutants we see that residues that are observed in the MSA are also detected in the saturation mutagenesis experiment and vice versa; in other words: residues not observed in the MSA were not observed experimentally.

FAD-binding domain of the vanillyl alcohol oxidase super-family

The vanillyl alcohol oxidase (VAO) flavoprotein family (FAD-O in SCOP database) is a large group of enzymes that catalyze a wide variety of oxidation-reduction reac-

Table III

Residues Observed at Positions in the MSA that Have At Least Four Out of Six Glucokinase Fingerprint Residues According to the Consensus, Together with the Results of Saturation Mutagenesis at Four of these Positions[22]

	Ali	Screen	Ali	Screen	Ali	Screen	Ali	Screen	Ali	Ali
No.	46		47		67		88		52	82
A	WT	WT	−	−	−	−	−	−	+	+
C	−	+	+	+	−	+	−	−	−	−
D	−	−	−	+	−	−	−	−	−	−
E	−	−	−	−	−	−	−	−	−	−
F	−	−	+	+	WT	WT	−	−	−	−
G	+	+	WT	WT	−	−	WT	WT	−	WT
H	−	−	+	+	−	−	−	−	−	−
I	−	−	+	+	+	+	−	−	−	−
K	−	−	−	+	−	+	−	−	+	−
L	−	−	+	+	+	+	−	−	−	−
M	−	−	+	+	+	+	−	−	−	−
N	−	−	+	−	−	+	−	−	WT	−
P	+	+	−	−	+	+	−	−	−	−
Q	−	−	−	−	−	+	−	−	−	−
R	−	−	−	+	−	−	−	−	−	−
S	+	+	+	+	+	+	−	−	−	−
T	−	−	+	+	+	+	−	−	−	−
V	−	−	+	+	+	+	−	−	−	−
W	−	−	−	−	−	+	−	−	−	−
Y	−	−	−	+	−	−	−	−	−	−

The residue numbers shown are 3DM-numbers. The corresponding numbers in the *E. coli* glucokinase are 64, 65, 101, 140, 76, 134, respectively. The "Ali" columns show a plus sign if the amino acid type was detected at least once in the MSA at that position. Plus signs in the "Screen" columns indicate that the saturation mutagenesis experiment showed that this residue type at that position produced viable protein. Minus signs indicate non-observed residue types. WT indicates that that residue either is the consensus residue at that position, or is observed in the *E. coli* wild type sequence. Not counting the WT cases, we observe agreement in 65 out of 76 cases.

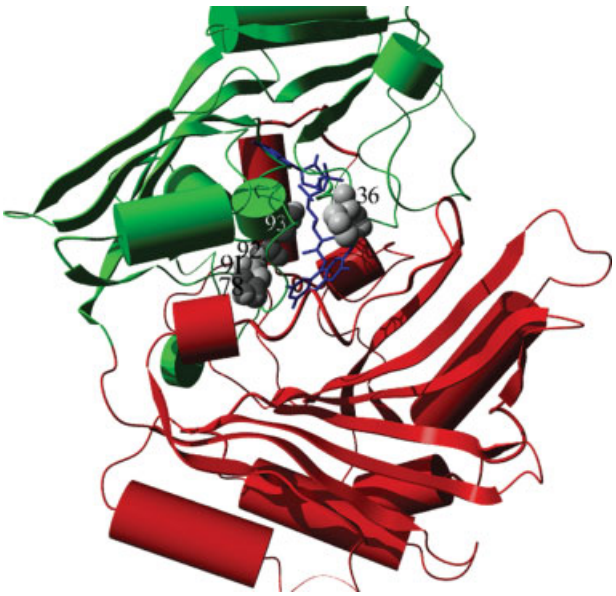

Figure 3

Crystal structure of 6-hydroxy-D-nicotine oxidase (PDB accession code: 2BVF). The FAD-binding domain is in green, the cap-domain is in red, the 8 α-N1-histidyl-FAD cofactor (in blue), and the residues with high CM scores (in gray) are shown as ball models.

tions.[23,24] Members of this family share a characteristic domain topology, with a conserved N-terminal FAD-binding domain and a less well conserved C-terminal cap domain that determines the substrate specificity (see Fig. 3). Most members of the VAO flavoprotein family contain a covalently bound FAD cofactor. L-Galactono-1,4-lactone dehydrogenase (GALDH; SwissProt ac = Q8GY16) is a VAO-family member that is involved in the vitamin C biosynthesis pathway in plants. Structural information is neither available for GALDH nor for any close homolog. Consequently little is known about GALDH's active site or about the molecular basis for the noncovalent binding of the FAD cofactor, and a series of mutations was therefore made to obtain such information. Position 56 (36 in the 3DM alignment) is located in the so-called PP-loop that interacts with the pyrophosphate moiety of the FAD molecule.[23] In most family members with a covalently bound FAD, a histidine is observed at this position. Replacing the histidine at position 36 in covalent VAO-family members yielded either active proteins with noncovalently bound FAD, or inactive apo-proteins.[25,26]

A 3DM alignment was constructed using 1152 sequences of (putative) VAO-family members. 3D protein structures are sparsely spread over this wide enzyme superfamily. The 3DM alignment of the VAO-flavoprotein super-family comprises only the N-terminal FAD-binding domain due to a lack of structural conservation in the C-terminal cap domain. Alignment position 36 (a histidine in the VAO members with a covalently bound FAD) correlates well with positions 78, 91, and 92. Positions 78, 91, and 92 are all located in the direct vicinity of the pyrophosphate moiety of the isoalloxazine ring of the flavin, with residue 78 being at hydrogen bonding distance of the reactive N5 locus (see Fig. 3).

Table IV

Catalytic and FAD-Binding Properties of GALDH Variants

Variant	k_{cat} (s^{-1})	K_m (m*M*)	FAD binding (F-factor)
Wild-type[a]	134 ± 5	0.17 ± 0.01	++ (8.0)
L36H[a]	32 ± 1	0.12 ± 0.01	++ (7.9)
A78G	116 ± 5	0.45 ± 0.03	++ (8.3)
V91G	57 ± 7	0.26 ± 0.02	± (14.1)
L36H/A78G	7.6 ± 0.2	0.15 ± 0.02	++ (8.2)
L36H/V91G	23 ± 2	0.16 ± 0.01	± (14.6)
A78G/V91G	49 ± 2	0.31 ± 0.06	+ (10.5)
L36H/A78G/V91G	<0.1	ND	—(ND)

[a]Adapted from Leferink *et al.*[24], ND is not determined.

Among the VAO-family members with a covalently bound FAD, a histidine is favored at position 36, and glycines are favored at positions 78, 91, and 92. MurB reductases, which have a noncovalently bound FAD, favor a serine at position 36, and a leucine or alanine at position 78 and a methionine at position 91. GALDH contains a leucine at position 36 (Leu56), an alanine at position 78 (Ala113), and a valine at position 91 (Val126). Throughout the super-family, glycine is the preferred residue type at position 92 both for variants with a His at position 36 and for variants with a Leu at this position. Mutations studies were therefore started with the single mutants L36H, A78G, and V91G. These are the three mutations that move the GALDH sequence in the direction of the consensus of family members with a covalently bound FAD. Covalently bound FAD was not observed in any of the GALDH variants.

Table IV shows that the three single mutants have similar or worse K_m and k_{cat} values than the wild-type enzyme. Table IV also shows that V91G poorly binds FAD. The Comulator residue type frequencies (Table V) indicate that all three single mutations led to situations with unfavorable amino acid combinations of the residue pairs shown to be important by the CMA.

With a His at position 36 the most abundant residue at position 78 is a Gly. When this Gly is added in the

Table V

Comulator Residue Type Frequencies of GALDH Variants

	36 78 91	36/78	36/91	78/91
WT	L A V	L/A (0.26)	L/V (0.09)	A/V (0.09)
L36H	H A V	H/A (0.35)	H/V (0.69)	A/V (0.09)
A78G	L G V	L/G (0.17)	L/V (0.09)	G/V (0.26)
V91G	L A G	L/A (0.26)	L/G (0.0)	A/G (0.61)
L36H/A78G	H G V	H/G (**19.9**)	H/V (0.69)	G/V (0.26)
L36H/V91G	H A G	H/A (0.35)	H/G (**15.9**)	A/G (0.61)
A78G/V91G	L G G	L/G (0.17)	L/G (0.0)	G/G (**15.4**)

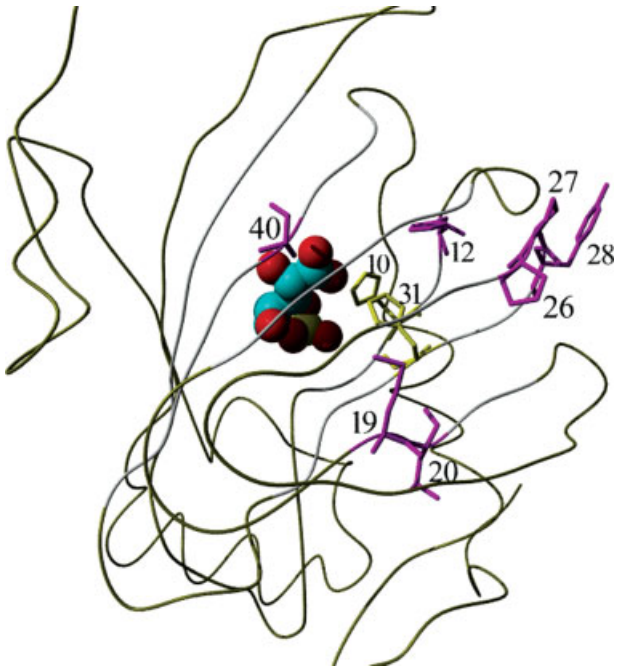

Figure 4

Tube representation of the 3D-structure of PfPGI from *P. furiosus* (PDB accession code: 1 × 82). The two conserved histidines are shown in yellow and the core positions with high CM scores are in magenta. The inhibitor 5-phospho-D-arabinonate is shown as ball model.

L36H/A78G double mutant, the catalytic rate is not improved, but the deteriorating effect on K_m of the A78G mutant is compensated (Table IV). Note that the loss in k_{cat} in the L36H mutant is much stronger in the A78G background than in the single mutant. With a His at position 36 the most abundant residue at position 91 is a Gly (Table V). When this Gly is added in the L36H/V91G double mutant, we see a similar effect as observed for the L36H/A78G double mutant, an improvement in the K_m but a loss of k_{cat}. With a Gly at position 91, the most abundant residue at position 78 is a Gly. When Gly78 was added in the A78G/V91G double mutant, a variant was obtained with catalytic properties comparable to both single mutants, but with much better FAD-binding properties than V91G (Table IV). Introducing a His at position 36 in the L36H/A78G/V91G triple mutant resulted in a mutant protein that is expressed as insoluble apo-protein that could not be purified in enough quantities to perform biochemical studies.

RmlC-like cupin super-family

The alignment of the cupin super-family is the largest of the four super-families studied and contains 2097 sequences. The RmlC-like cupin super-family consists of proteins possessing a common β-barrel structure also known as a jelly roll fold. Although the proteins in this super-family are functionally diverse,[27] most are enzymes of which the active site is located within the β-barrel. This active site often contains two histidines (3DM-numbers 10 and 31) that are conserved in ~80% of all sequences (Fig. 4: yellow residues). Comulator revealed a network of highly correlating positions consisting of the alignment positions with 3DM-numbers 12, 19, 20, 26, 27,28, and 40 (Fig. 4 magenta residues). The highest pair-wise CM score was detected for position pair 27 and 28. The alignment positions 26–28 form a structurally conserved surface loop in most members of the super-family.

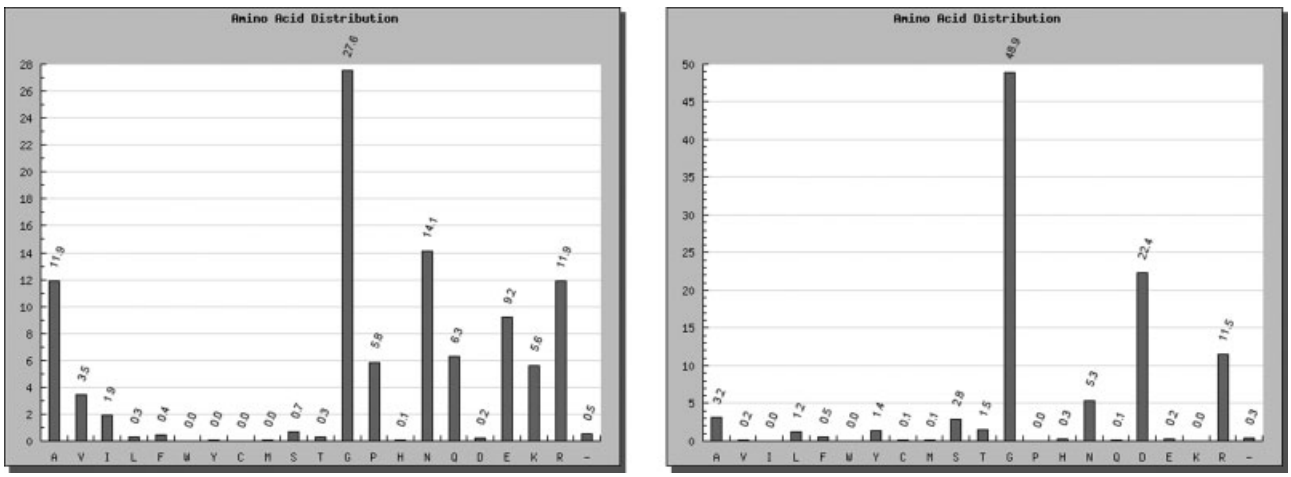

Figure 5

Bar graphs representing the amino acid distributions of positions 27 (left) and 28 (right) of the RmlC-like cupin super-family. The *x*-axis lists the 20 different amino acids and the *y*-axis their percentages in the MSA.

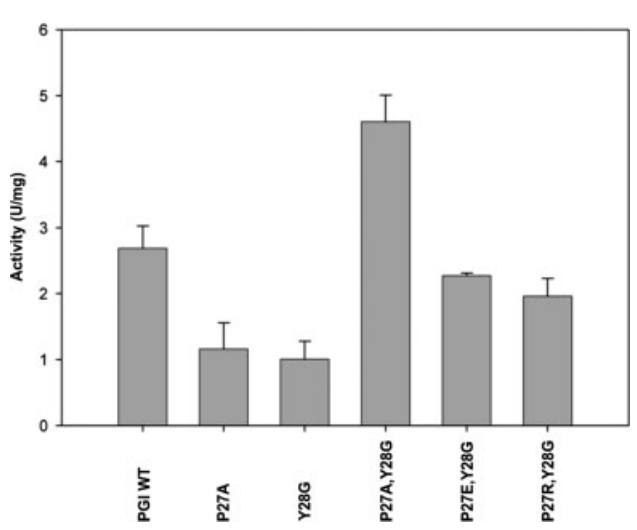

Figure 6

Bar graphs representing activity of single or double mutants of wild type PfPGI. Numbering according to the 3DM numbering. The numbers in the amino acid sequence of PfPGI are 132 and 133, respectively.

The 3DM software automatically extracted many hundreds of mutations for the cupin super-family from the literature and stored them in the database to allow for rapid inspection. Intriguingly, the network of correlated positions has barely been mutated. Mutagenesis of position 28 in flavonol synthase from *Citrus unshiu* (G261A) resulted in 95% reduction of enzyme activity. Introduction of a proline at this same position 28 resulted in a completely inactive enzyme.[28] The residue at position 28 is located far away from the active site histidines (see Fig. 4) but nevertheless has been shown important for activity. Because the residue positions 27 and 28 show the highest CMA value, we decided to study this pair of residues experimentally.

One of the best characterized members of the cupin superfamily is the *Pyrococcus furiosus* PGI (PfPGI).[18,29,30] Several crystal structures of this protein have been elucidated[18,29–31] and the reaction mechanism has been analyzed by mutagenesis, NMR, and EPR studies.[30] PfPGI has a tyrosine at position 28 whereas glycine is the most prevalent amino acid at this position in the cupin super-family (42%, see Fig. 5). The Y28G mutant (Y133G in PfPGI numbering) results in a 2.6-fold reduction of the activity (see Fig. 6). The most abundant residue at position 27 (that is highly correlated with 28) is glycine (present in 26% of all sequences); in PfPGI it is a proline residue. Analysis of the Comulator heat-map reveals that the combination P27-Y28 occurs rarely (1.1%), whereas P27-G28 occurs in 4.4% of all sequences. However, when a glycine is observed at position 28, an alanine is by far the most prevalent residue at position 27 (23%). The double mutation P27A/Y28G not only regained the activity lost by the Y28G mutant, but even became twice as active as the wild-type enzyme. Obviously, if we had started with the P27A mutant and compensated it with Y28G, we would have obtained the same result (see Fig. 6).

Of the sequences with a glycine at position 28, 18.5% has a glutamate at position 27, and 18.2% an arginine. Both P27E and P27R can compensate for the loss of activity of Y28G regaining near WT activity (see Fig. 6).

CONCLUSIONS

Most enzyme engineering successes of the past decade have been accomplished via random mutagenesis, euphemistically called evolutionary approaches, while rational mutagenesis approaches in terms of predicting one mutation at one position to achieve one phenotypic effect were often less successful. The recent explosion in a series of high-throughput technologies, including sequencing, is enabling an even larger speed in the technical execution of these evolutionary approaches. It has often been observed, and this study adds one more observation, that optimal phenotypic effects tend to require a series of mutations to be introduced simultaneously, and many evolutionary approaches are optimized to just achieve that goal. Still, parallel random mutagenesis is technically limited to a handful of amino acid positions in the protein. It is therefore of paramount importance to select the positions well where these random mutations are going to be introduced. We have applied the Comulator tool to 10 enzyme families. In all the cases the highly correlating positions are observed surrounding the active site. The results for the four families for which mutation information was available for at least one of those positions is reported. The other six families, however, all at least support the idea that correlating positions are related to functional rather than structural aspects. Thus with CMA we can find groups of residues that are involved in the same function. And we have shown that this enables us to find the combinations of mutations that improve catalysis rate or modify substrate specificity. It seems therefore that a major step forward can be made in enzyme engineering if the amino acid positions selected for combined randomization are carefully selected from a CMA screen.

Our results also show that the combinatorial freedom at the positions detected by the CMA is limited so that full randomization is not needed to harvest the complete combinatorial potential. Our mutation studies have shown that the combination of residues that can bring the desired phenotypic change in the enzyme often has already been tried in a different context, that is, in another protein, so that the limited number of sequence fingerprints obtained by CMA will be a good start for limited randomization. Looking at the ease of today's

gene synthesis approaches, we can imagine that this might be a new path towards semi rational enzyme engineering.

REFERENCES

1. Halperin I, Wolfson H, Nussinov R. Correlated mutations: advances and limitations. A study on fusion proteins and on the Cohesin-Dockerin families. Proteins 2006;63:832–845.
2. Fodor AA, Aldrich RW. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. Proteins 2004;56:211–221.
3. Pazos F, Helmer-Citterich M, Ausiello G, Valencia A. Correlated mutations contain information about protein-protein interaction. J Mol Biol 1997;271:511–523.
4. Mintseris J, Weng Z. Structure, function, and evolution of transient and obligate protein-protein interactions. Proc Natl Acad Sci USA 2005;102:10930–10935.
5. Kundrotas PJ, Alexov EG. Predicting residue contacts using pragmatic correlated mutations method: reducing the false positives. BMC Bioinformatics 2006;7:503.
6. Rost B. Sean O'Donoghue Sisyphus and protein structure prediction. Bioinformatics 1997;13:345–356.
7. Oliveira L. Paiva A. Vriend G. J. A model for G-protein coupled receptors. Comp Aided Mol Des 1993;7:649–658.
8. Singer MS, Oliveira L, Vriend G, Shepherd GM. Potential ligand-binding residues in rat olfactory receptors identified by correlated mutation analysis. Receptors Channels 1995;3:89–95.
9. LinksPulim V, Bienkowska J, Berger B. LTHREADER: prediction of extracellular ligand-receptor interactions in cytokines using localized threading. Protein Sci 2008;17:279–292.
10. Garriga C, Pérez-Elías MJ, Delgado R, Ruiz L, Nájera R, Pumarola T, Alonso-Socas Mdel M, García-Bujalance S, Menéndez-Arias L. Mutational patterns and correlated amino acid substitutions in the HIV-1 protease after virological failure to nelfinavir- and lopinavir/ritonavir-based treatments. J Med Virol 2007;79:1617–1628.
11. Liu Y, Eyal E, Bahar I. Analysis of correlated mutations in HIV-1 protease using spectral clustering. Bioinformatics 2008;24:1243–1250.
12. Hubbard TJ, Ailey B, Brenner SE, Murzin AG, Chothia C. SCOP: a structural classification of proteins database. Nucleic Acids Res 1999:27;254–256.
13. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
14. Shulman AI, Larson C, Mangelsdorf DJ, Ranganathan R. Structural determinants of allosteric ligand activation in RXR heterodimers. Cell 2004;116:417–429.
15. Lockless SW, Ranganathan R. Evolutionarily conserved pathways of energetic connectivity in protein families. Science 1999;286:295–259.
16. Fodor AA, Aldrich RW. On evolutionary conservation of thermodynamic coupling in proteins. J Biol Chem 2004;279:19046–19050.
17. Verhees CH, Huynen MA, Ward DE, Schiltz E, de Vos WM, van der Oost J. The phosphoglucose isomerase from the hyperthermophilic archaeon *Pyrococcus furiosus* is a unique glycolytic enzyme that belongs to the cupin super-family. J Biol Chem 2001;276:40926–40932.
18. Berrisford JM, Akerboom J, Turnbull AP, de Geus D, Sedelnikova SE, Staton I, McLeod CW, Verhees CH, van der Oost J, Rice DW, Baker PJ. Crystal structure of *Pyrococcus furiosus* phosphoglucose isomerase. Implications for substrate binding and catalysis. J Biol Chem 2003;278:33290–33297.
19. Leferink NGH, Van den Berg WAM, Van Berkel WJH. L-Galactono-γ-lactone dehydrogenase from *Arabidopsis thaliana*, a flavoprotein involved in vitamin C biosynthesis. FEBS J 2008;275:713–726.
20. Joosten HJ, Han Y, Niu W, Du J, Vervoort J, Dunaway-Mariano D, Schaap PJ. Identification of fungal oxaloacetate hydrolyase within the isocitrate lyase/PEP mutase enzyme super-family using a sequence marker based method. Proteins 2008;70:157–166.
21. Narayanan B, Niu W, Joosten HJ, Kuipers RKP, Li Z, Schaap PJ, Dunaway-Mariano D, Herzberg O. Structure and function of 2,3-dimethylmalate lyase, a PEP Mutase/isocitrate lyase superfamily member. J Mol Biol 2009;386:486–503.
22. Miller BG. The mutability of enzyme active-site shape determinants. Protein Sci 2007;16:1965–1968.
23. Fraaije MW, Van Berkel WJ, Benen JA, Visser J, Mattevi A. A novel oxidoreductase family sharing a conserved FAD-binding domain. Trends Biochem Sci 1998;23:206–207.
24. Leferink NGH, Heuts DPHM, Fraaije MW, Van Berkel WJ. The growing VAO flavoprotein family. Arch Biochem Biophys 2008;474:292–301.
25. Caldinelli L, Iametti S, Barbiroli A, Fessas D, Bonomi F, Piubelli L, Molla G, Pollegioni L. Relevance of the flavin binding to the stability and folding of engineered cholesterol oxidase containing noncovalently bound FAD. Protein Sci 2008;17:409–419.
26. Heuts DP, van Hellemond EW, Janssen DB, Fraaije MW. Discovery, characterization and kinetic analysis of an alditol oxidase from *Streptomyces coelicolor*. J Biol Chem 2007;282:20283–20291.
27. Dunwell JM, Purvis A, Khuri S. Cupins: the most functionally diverse protein super-family. Phytochemistry 2004:65;7–17.
28. Wellmann F, Lukacin R, Moriguchi T, Britsch L, Schiltz E, Matern U. Functional expression and mutational analysis of flavonol synthase from *Citrus unshiu*. Eur J Biochem 2002;269:4134–4142.
29. Berrisford JM, Akerboom J, Brouns S, Sedelnikova SE, Turnbull AP, van der Oost J, Salmon L, Hardré R, Murray IA, Blackburn GM, Rice DW, Baker PJ. The structures of inhibitor complexes of *Pyrococcus furiosus* phosphoglucose isomerase provide insights into substrate binding and catalysis. J Mol Biol 2004;343:649–657.
30. Berrisford JM, Hounslow AM, Akerboom J, Hagen WR, Brouns SJ, van der Oost J, Murray IA, Michael Blackburn G, Waltho JP, Rice DW, Baker PJ. Evidence supporting a cis-enediol-based mechanism for *Pyrococcus furiosus* phosphoglucose isomerase. J Mol Biol 2006;358:1353–1366.
31. Hansen T, Oehlmann M, Schonheit P. Novel type of glucose-6-phosphate isomerase in the hyperthermophilic archaeon *Pyrococcus furiosus*. J Bacteriol 2001;183:3428–3435.