Electronic Supplementary Material (ESI) for Molecular BioSystems
This journal is © The Royal Society of Chemistry 2012

Supporting text

Cascaded walks in protein sequence space: Use of artificial sequences in remote homology detection between natural proteins

S. Sandhya, R. Mudgal, C. Jayadev, K.R. Abhinandan, R. Sowdhamini and N. Srinivasan

S1) Choice of level: a handle on extent of similarity between designed and natural sequences:

The choice of residues for the twenty standard amino acid residues for each position in the PSSM to the first 'k' residue types was introduced through the selection of a 'level' for designing the sequence. This feature introduces a handle on the extent of allowed sequence changes permitted at each alignment position in the algorithm. For instance, a choice of level two would strictly shortlist only the top two most frequently occurring residue types in every alignment position of the profile. It is anticipated therefore, that at level 2, the resultant artificially generated sequences are at least 80% similar to their parent sequences at each alignment position. Level 20 on the other hand, would permit choice of all the twenty residue types observed at that position. At this level, the sequences are likely to be most distant from the parent sequence. At strictly conserved positions where only a specific amino-acid is preferred, level 20 of sequence design would not select for any of the twenty residues since the top 20 residues at that position would always be the most conserved/preferred residue. This sort of restriction is essential to retain family-specific features and variation is introduced primarily in less conserved regions in the alignment.

S2) Assessments of designed sequences at different levels

To assess the influence of level on the quality of designed sequences for a domain family, artificially generated sequences were obtained at different levels of design (ranging from 1 -20). Each generated sequence was subjected to an RPS-BLAST search against a database of the 40,000 parent profiles/ PSSMs of the entire PALI dataset. A criteria that each biased random sequence would identify its parent family as the best hit at an E-value better than 0.0001 was imposed for consideration as a 'designed sequence' and only such sequences were used in further assessments. The number of such generated sequences in each level that qualified the goodness criteria was noted. Appropriate level was selected when at least 50% of the biased random sequences generated for a family at a particular level were able to identify the cognate parent family profile at an E-value better than 0.0001. Since a large number of sequences could be generated for a domain family through the roulette wheel-based approach, we generated twice the number of natural members of the protein family to populate the database.

S3) Generating artificial sequences at various levels

This journal is © The Royal Society of Chemistry 2012

As seen in Figure S1, for a single domain family the number of sequences that detect their parent family profiles at appreciable E-value (<0.0001) at different levels (Levels 2, 5 10, 15 and 20) varies. Even at level 20, some of the generated sequences are able to detect their parent profiles, although a small percentage. For each family therefore, depending on the number of sequences that qualify, a different level of design was chosen. Care was taken to ensure that at least 50% of the generated sequences would qualify the selection criteria for designed sequences.

The choice of an optimal 'k' (or level) for a protein domain family could depend on the inherent divergence and sequence features of the family. Close knit-families such as the chorismate lyase, S-adenosylmethionine carboxylase, Fe-only hydrogenase folds are less sequence divergent than other expansive folds such as the globins and the P-loop NTP hydrolases. In such domain families, choice of levels two to five of design would generate sequences very similar to the parent members and such sequences in the database would bias the database searches towards the parent family and restrict sequence hops across distantly related members. Therefore, in such cases the need for more divergent sequences to populate the database may favor the choice of a more relaxed level during the design process.

S4) Types of PSSMs employed in sequence design

PSSMs can be obtained from a multiple sequence alignment of a protein family. This may be done by jump-starting the PSI-BLAST search with an alignment and selecting any one reference query from the family (1). To query, a null database with no sequence data is employed so that only information from the input alignment is employed in generating the PSSM after two iterations of the search. An E-value of 0.0001 was employed in the PSI-BLAST run. Two types of PSSMs were employed to represent a protein family. In the first type, a single representative PSSM was derived for a protein family, using any one reference query sequence from the protein family. In principle such a profile should capture the sequence features of all members of the domain family. Single PSSMs were thus generated for each PALI family alignment and each profile was labeled as an SPSSM-derived profile. A total of 1444 such PSSMs were generated for each of the 1444 PALI families. However, protein sequence space even within a domain family is observed to be non-uniform and sub-clustering of sequences due to variation in sequence identities across domain family members is observed in families such as the IgG-like, P-loop NTP hydrolase, Lipocalin and other folds (Figure S2a) (2). In the second type of PSSM generated for a domain family, multiple reference queries were used to derive PSSMs for the family resulting in more than one PSSM per family (MPSSM/ multiple PSSM). The problem of sequence dispersion and representation in PSSM is therefore very carefully addressed and every domain family member finds membership in at least one of the many PSSMs generated for the domain family. Earlier studies involving multiple PSSMs

This journal is © The Royal Society of Chemistry 2012

have shown that they were more effective than using traditional single PSSMs and Hidden Markov Models (3). Multiple PSSMs corresponding to each PALI domain family (40,946 PSSMs, representing 1444 protein families), already available in the MulPSSM database (4), were directly employed in the assessment. The multiple PSSMs associated with a domain family were tagged with identifiers/ SCOP codes of the parent domain family to keep track of the associations at the time of analysis.

S5) Generating designed sequences from PSSMs

a) 'Tagging' of PSSM-derived sequences

Each PSSM was employed to generate artificial sequences using the roulette wheel approach (Fig 1). Since both SPSSM and MPSSM-derived PSSMs were tested, the artificially generated sequences from either type of PSSM were annotated differently. Where an SPSSM of a SCOP family was employed as the parent profile, the generated sequence inherited the SCOP code of the parent family in its description line. When multiple PSSMs were employed, the generated sequences inherited, in addition to the SCOP code of the parent family, an identifier for the PSSM from which they were derived. This identifier was retained to facilitate associations of the generated sequences to the specific sub-cluster of proteins within a protein family from which they were derived.

S6) How representative of the family are sequences derived from SPSSMs and MPSSMs

Using the roulette-wheel approach (Fig 1), 'designed sequences' were generated for each domain family at the appropriate level. Since the number of designed sequences generated for each domain family was twice the number of actual domain members of a family, care was taken to ensure that this number was equally divided across each of the MPSSMs. To describe the quality and similarity of designed sequences with respect to parent domain family members, SPSSM and MPSSM-derived sequences were merged with their cognate family members separately. Multiple alignments of sets of such sequences were subjected to phylogenetic analysis (Fig. S2b, Fig. 3c).

S7) Parameter tuning

The following parameters were tested in the sequence design procedure. These include:

a) Choice of E-value at which to perform searches

Searches were performed in the augmented database at E-values ranging from 0.0001 to 1.0 for different queries from the TIM, cytochrome, lipocalin, OB, DNA-glycosylase and P-loop NTP hydrolase folds. The number of fold members recognized at the end of three generations was

This journal is © The Royal Society of Chemistry 2012

determined for each of the E-values. The performance at different E-values was measured through computations of specificity, sensitivity and determination of error rates. For the purpose of assessment, the following definitions have been employed:

True positives: Number of fold members detected by the search protocol (at the end of three generations of Cascade PSI-BLAST search).

True negatives: Number of non-fold members identified as belonging to another fold by the search method.

False positives: Number of non-fold members detected as hits by the search method.

False negatives: Number of true fold members that remained undetected by the search method.

Specificity: Specificity is a measure of the ability to distinguish true members from the total number of false entries in the dataset. Specificity= TN/FP+TN

FPR= (1-Specificity)

Sensitivity: Sensitivity is a measure of the number of true positives detected as a fraction of the total number of true entries in the dataset. Sensitivity = TPR = TP/ (TP+FN)

Receiver operating characteristics curves (ROC) that compare the tradeoffs between sensitivity and false positive rates, at various E-value thresholds were also determined.

Lower E-values favor fewer false positives (Fig. S3). Higher E-value thresholds (>0.01) show improvements in sensitivity but also detect more false-positives. To minimize the occurrence of false positives, a stringent threshold of 0.0001 is therefore, applied in all the searches described here.

This journal is © The Royal Society of Chemistry 2012

b) Choice of alignment length at which to design sequences

Cascade PSI-BLAST searches were also performed using different alignment length filters. Different alignment length thresholds of 60, 45 and 30% were tested for members of the P-loop NTP hydrolase fold that show nearly two-fold domain length differences. The influence of this parameter on coverage is seen to be case-specific and ROC curves at different length cut-offs show that some domain families, inherently variable in length such as P-loop NTP hydrolase, respond to relaxations in length-filters better than other domain families that are length-rigid (5).

References

1. Altschul SF, *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25(17):3389-3402.
2. Salamov AA, Suwa M, Orengo CA, Swindells MB (1999) Combining sensitive database searches with multiple intermediates to detect distant homologues. *Protein Eng* 12(2):95-100.
3. Anand B, Gowri VS, Srinivasan N (2005) Use of multiple profiles corresponding to a sequence alignment enables effective detection of remote homologues. *Bioinformatics* 21(12):2821-2826.
4. Gowri VS, Sandhya S (2006) Recent trends in remote homology detection: an Indian Medley. *Bioinformation* 1(3):94-96.
5. Sandhya S, *et al.* (2009) Length variations amongst protein domain superfamilies and consequences on structure and function. *PLoS One* 4(3):e4981.

This journal is © The Royal Society of Chemistry 2012

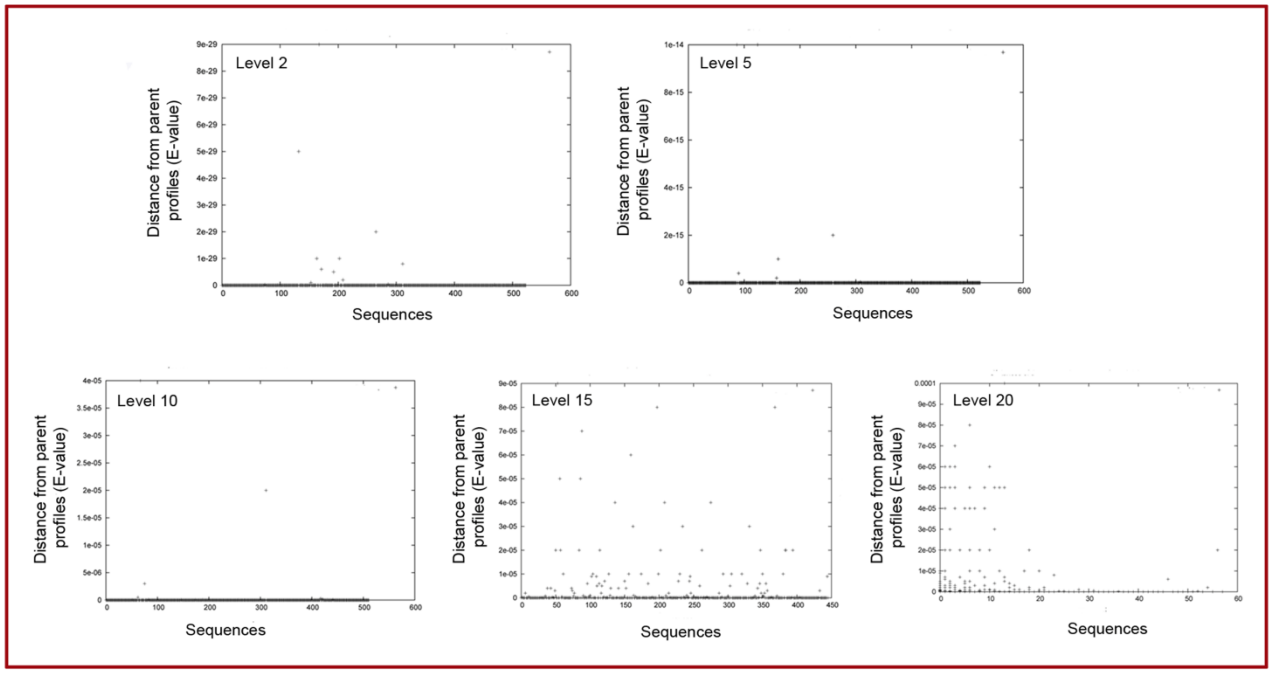

Fig. S1. Distribution of the generated sequences at different Levels: Plotted on X-axis are the generated sequences for the Lipocalin fold. On Y-axis, the distance of the generated sequences from the parent profiles (as E-values) is plotted.

This journal is © The Royal Society of Chemistry 2012

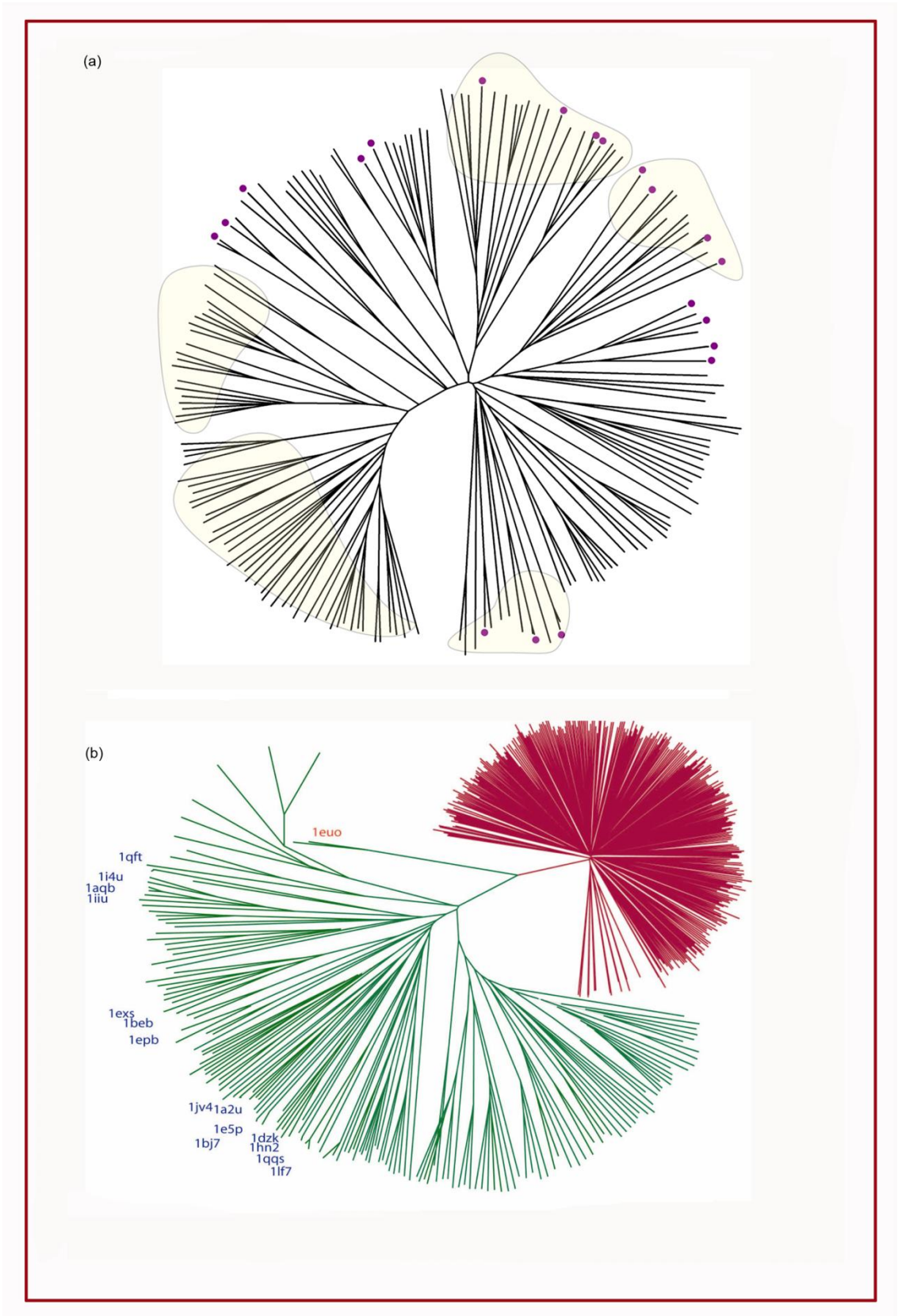

Fig. S2. Sequence dispersion in the lipocalin fold:
a) The domain family members of the Lipocalin fold group into distinct sub-clusters (encircled in yellow). Domain members with structures are indicated by purple dots.
b) SPSSM-derived designed sequences (in red) are biased to parent query (1euo, in red) and unrepresentative of all domain family members (in green).

This journal is © The Royal Society of Chemistry 2012

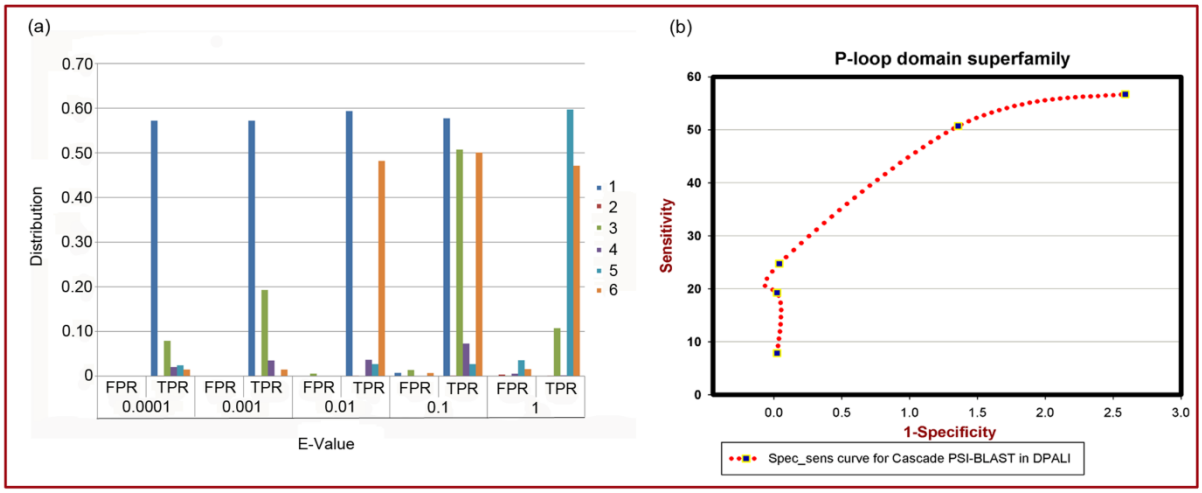

Fig. S3. Performance of designed sequences in Cascade PSI-BLAST searches.
a)Distribution of False positives and true positives at different E-values.
b)ROC curves for P-loop NTP-hydrolase fold members at different E-value thresholds.

This journal is © The Royal Society of Chemistry 2012

Table S1: Coverage of protein family through PSI-BLAST and Cascade PSI-BLAST searches in DPALI and PALI+ databases. (Numbers enclosed in parenthesis that follow the protein description are the number of known structural members of the family).

S.No	Protein name	Query	Number of Fold members	True- positives in DPALI search			True- positives in PALI+ search			Family-coverage in Cascade PSI-BLAST searches (in %)		% Coverage of structural members of the family (Cascade PSI-BLAST)	
				I	II	III	I	II	III	DPALI	PALI+	DPALI	PALI+
1	Retinol-binding protein (20)	**1KT7**	**123**	105	9	5	112	2	-	96.7	93	95	85
2	Odorant-binding protein (20)	**1HN2**	**123**	111	7	-	101	-	-	96	82	95	85
3	Lipocalin allergen (20)	**1BJ7**	**123**	111	3	5	114	-	-	96.7	93	95	85
4	Salivary lipocalin (20)	**1GM6**	**123**	109	10	-	112	2	-	96.7	93	95	85
5	Aphrodisin (20)	**1E5P**	**123**	109	10	-	112	2	-	96.7	93	95	85
6	beta-Lactoglobulin (20)	**1BEB**	**123**	112	7	-	111	3	-	96.7	93	95	85
7	Retinoic acid-binding protein (20)	**1EPB**	**123**	110	8	-	114	-	-	96	93	95	85
8	Major urinary protein (20)	**1JV4**	**123**	111	8	-	114	-	-	96.7	93	95	85
9	Neutrophil gelatinase-associated lipocalin (20)	**1QQS**	**123**	107	12	-	112	2	-	96.7	93	95	85
10	Lipocalin q83 (20)	**1JZU**	**123**	112	7	-	114	-	-	96.7	93	95	85
11	Complement protein C8gamma (20)	**1LF7**	**123**	102	17	-	110	4	-	96.7	93	95	85
12	Bilin-binding protein (20)	**1KXO**	**123**	104	15	-	82	32	-	96.7	93	95	85
13	Alpha-crustacyanin (20)	**1I4U**	**123**	103	16	-	25	89	-	96.7	93	95	85
14	Nitrophorin 1 (20)	**1NP1**	**123**	5	108	6	5	-	-	96.7	4	95	1
15	Nitrophorin 2 (20)	**1EUO**	**123**	6	107	7	5	-	-	97.6	4	95	1
16	Cellular retinoic-acid-binding protein (19)	**1CBS**	**82**	81	-	-	80	1	-	99	99	100	100
17	Viral structural mimic of eIF2alpha (18)	**1LUZ**	**183**	22	9	8	2	-	-	21.3	1	11.1	5.5
18	C-terminal domain of RNA polymerase II subunit RBP4 (18)	**1GO3**	**183**	2	26	3	2	19	5	17	14.2	11.1	5.5
19	Ribosomal protein S17 (18)	**1HR0**	**183**	6	-	-	5	-	-	3	2.7	5.5	5.5
20	C-terminal domain of metazoan tyrosyl-tRNA synthetase (4)	**1NTG**	**49**	1	18	30	1	-	-	100	2	50	25
21	EMAP II (4)	**1FL0**	**49**	1	40	8	1	-	-	100	2	100	25
22	RNA guanylyltransferase (3)	**1CKM**	**39**	5	6	-	1	-	-	28.2	2.5	-	-
23	mRNA capping enzyme alpha	**1P16**	**39**	3	7	1	3	7	-	28.2	25.6	-	-

This journal is © The Royal Society of Chemistry 2012

	subunit (3)												
24	Ascaris hemoglobin, domain 1	**1ASH**	**133**	124	7	-	8	118	5	98.5	98.5		
25	Dehaloperoxidase	**1EW6**	**133**	124	3	-	3	104	11	95.5	88.7		