Supplementary information
ASPD (Artificially Selected Proteins/Peptides Database) – a database of proteins and peptides evolved in vitro.

How to query ASPD using SRS system.

First, click the hyperlink SRS ACCESS: ASPD_ALIGN on the title page of the ASPD database. This way you access the ASPD_ALIGN table, installed under the SRS system, of the ASPD database.

Second, click "Search" on the SRS title page of the ASPD_ALIGN. This way you proceed to the query form .You can also switch to the extended query form by clicking the button "Extended query form".

Third, form your query and define the output view (which fields to display) on the query page and click "Submit query".

Examples of the possible queries:

If you want to find what proteins with known 3 dimensional structure, involved in the signal transduction, were studied by phage display. Then form the following query: "Keywords: signal; transduction; Link: PDB".

If you want to find how protease inhibitors were studied with phage display. Then try such query: "Template: protease inhibitor"

If you want to find substrates for certain enzyme, selected by the phage display (for example, protein-tyrosine kinase), you may try either the query "Target: Protein-tyrosine kinase", or, if you know the EC number for protein-tyrosine kinase (which is 2.7.1.112), you may try "Link: Enzyme 2.7.1.112".

If you want to find peptides retrieved by panning against anti-beta endorphin mAb, we suggest that you make the query "endorphin" in the field "Target", because the whole expression "anti-beta endorphin mAb" is unlikely to match exactly the description of this antibody (actually, in ASPD in this case we have " monoclonal antibody 3-E7 specific for N-terminus of beta-endorphin").

These are just the simplest examples and many other queries are possible.

Examples of the output:

According to how you set the view option, as the result of the search you will get the list of clickable identifiers, complete entries, or a selection of fields. If you select the option "Use predefined view: names only", you will get a list of entry identifiers (which do not tell you much about the content of the entry). If you select the option "Use predefined view: complete entries", you will get a list of complete entries, which in case the entries are numerous may take a while to download and may look cumbersome. The best choice in our opinion is to select fields to display from the menu "Create your own view Select fields to display: ". Here is an example of output, where the fields "Identifier", "Lit_reference", "Target", "Template" and "Link" are selected.

Queries based on literature references.

First, click the hyperlink SRS ACCESS: ASPD_REF on the title page of the ASPD database. This way you access the ASPD_REF table, installed under the SRS system, of the ASPD database.

Second, click "Search" on the SRS title page of the ASPD_REF. This way you proceed to the query form.

Third, form your query and define the view on the appearing query page and click "Submit query".

This way you will retrieve the full literary references. In order to proceed to the phage display data from these papers, you need to perform a search of the ASPD_ALIGN table using the field "Lit_reference" and the identifier of the entry from ASPD_REF as the query.

Research based on data stored in ASPD.

Pairwise correlation analysis.

We applied CRASP package [1] to calculate such pairwise correlations for all aligned sets of proteins and peptides

Let us consider a set of aligned amino acid sequences. Each sequence can be represented as a string of numerical values of a given amino acid property (hydrophobicity, volume etc.).

Let f_{ki} be the value of the property f at the i-th position of the k-th sequence, then the mean value of f in the i-column is $\overline{f}_i = \frac{1}{N^*} \sum_{k=1}^{N^*} f_{ki}$, where N^* is the number of sequences in the set, which have any amino acid (not gap) in the i-th column of the alignment.

The value of the linear correlation coefficient for the property f in the pair of columns i, j is: $r_{ij} = \frac{\sum_{k=1}^{N^*} (f_{ki} - \overline{f}_i)(f_{kj} - \overline{f}_j)}{\left(\sum_{k=1}^{N^*} (f_{ki} - \overline{f}_i)^2 \cdot \sum_{k=1}^{N^*} (f_{kj} - \overline{f}_j)^2 \right)^{1/2}}$.

To estimate the significance of correlation coefficients, amino acids were independently and randomly permuted within the columns of the original alignment. As a result, a set with the positional amino acid compositions identical to those in the original alignment but lacking any possible statistical relationship between the alignment positions was generated. Such "randomised" sets allow to estimate the significance level P for the correlation coefficient r_{ij}. This value was estimated as P=1-$m(|r_{ij}\ (\text{rand})| > |r_{ij}|)/M$, where $m(|r_{ij}\ (\text{rand})| > |r_{ij}|)$ is the number of randomised samples in which the correlation coefficient exceeded in its absolute value the r_{ij} value obtained for original sample, M is the total number of randomised samples. In this work, we have generated M=10^5 random alignments.

Each entry in the ASPD database was analyzed for presence of pairwise correlations in terms of 4 amino acid properties: volume [2], hydrophobicity [3], isoelectric point value [4] and polarity [5] and has a link entitled “Correlation analysis” to the pre-calculated results of such analysis. Below we consider an example of observed correlations.

We take immunoglobulin binding domain of peptostreptococcal protein L (PDB entry 2PTL, ASPD entries PH1PO007, PH1PO008, PH1PO009, PH1PO105, PH1PO106), selected in [6, 7] for immunoglobulin binding, which implies correct folding, for residues directly involved in binding were not randomized. The significant correlations (P >= 99%) are shown in Table 1 (hyperlinks are to the correlation matrices at the ASPD Web-site, which can be accessed from the original ASPD entries by the link “Correlation analysis”) and Figure 1.

Table 1. Pairs of positions in of the immunoglobulin binding domain of peptostreptococcal protein L (PDB entry 2PTL) showing significant (P >= 99%) correlations.

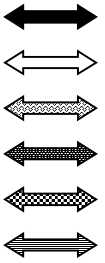
Phyico-chemical property	Alignment positions	Correlation value, r_{ij}	ASPD ID
Isoelectric point	K21-A34	-0.43	PH1BS029
	F26-S30	-0.53	PH1BS029
	F26-V18	-0.59	PH1PO105
	G38-A43	+0.85	PH1PO007
	T39-Y48	-0.79	PH1PO007
	I74-K75	+0.52	PH1PO009
Volume	F26-G29	-0.50	PH1PO106
Hydrophobicity	I20-K21	-0.52	PH1BS029
	V18-A22	-0.58	PH1PO105
	A66-D67	+0.54	PH1PO008

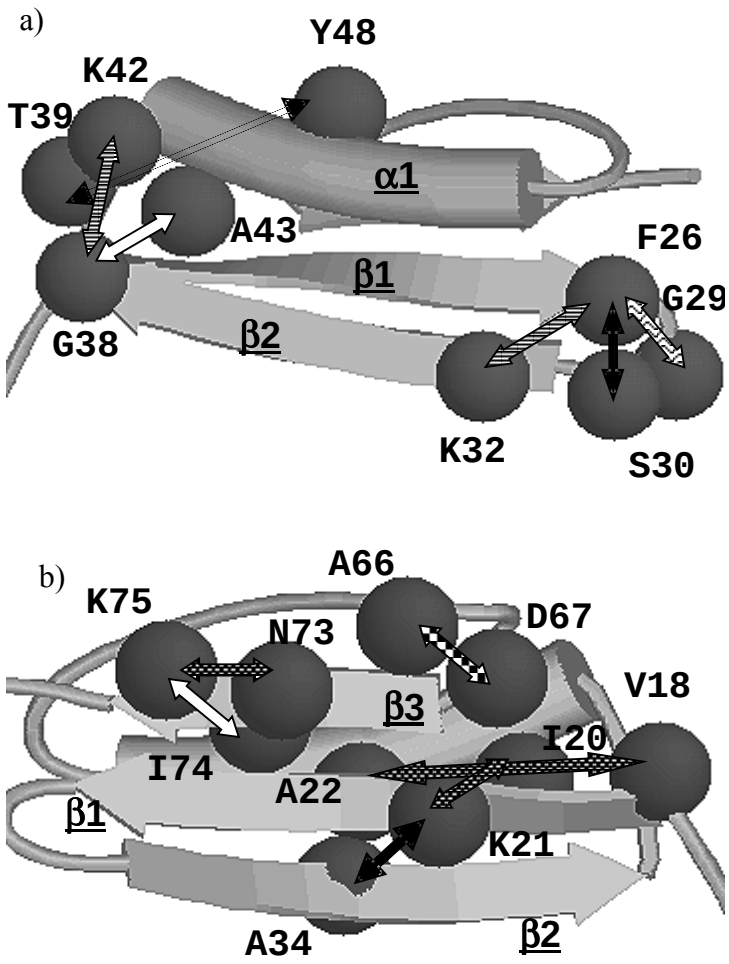
	N73-K75	-0.55	PH1PO009
Polarity	F26-Q32	+0.47	PH1BS029
	G38-K42	+0.375	PH1PO007

Pairs of positions A34-K21 and F26-S30 correlate negatively in the isoelectric point value. The first pair of positions lie adjacent on the C- and N- termini of antiparallel beta-sheets correspondingly, and the second pair on the same beta-sheets close to the turn. So we may suppose that electrostatic interaction of aminoacids in these positions are crucial in the beta-sheet formation. At the same time polarity of positions F26 and Q32 show positive correlation, which means a repulsion between these positions. One more residue that belongs to this correlating cluster is G29 which is situated exactly at the turn and it shows negative volume correlations with F26. So we observe a cluster of 4 residues (F26, S30, Q32 and G29) and their specific interactions (which we see as correlations) that is responsible for the correct beta-turn formation. Another cluster of correlating positions is found at the turn between the second beta-sheet and the alpha-helix (G38-A43, G38-K42). It is noteworthy, that it also includes a residue distant from this turn (correlation T39-Y48), like in previous case (K32). And the turn between the alpha-helix and the third beta-sheet also has a pair of correlating positions (A66-D67). There is also a negative correlation in hydrophobicity between positions I20 and K21. The residue I20 is situated inside the hydrophobic, and the pair of adjacent residues V18-A22, which have negative correlation in hydrophobicity, is also buried. Maybe these residues are responsible for hydrophobic core formation during its folding. The correlations within the cluster may be of similar nature (the residue I74 is buried and two others are exposed).

So we have shown, that the phage display technique, providing for great flexibility of the sequence, can yield data, the analysis of which in terms of correlations of amino acid properties provides insight into the mechanisms providing for structural integrity of protein, but one must look out for those dependencies which are connected with the experimental system.

Figure 1. Positions of residues that form significantly correlating pairs (a) in the turn regions (b) within the β-sheet in the structure of IgG binding protein L (PDB entry 2PTL) revealed by analysis of sequences stored in the ASPD. Correlations are indicated with arrows. Greek letters denote elements of the secondary structure.





Amino acid composition of the ASPD database.

The amino acid composition of the in vitro evolved proteins (only those amino acids which were retrieved via evolution were taken into account, those positions which had not been randomized were ignored) compared to that of SwissProt and to that of all active sites of proteins with known 3-dimensional structure (annotated as such in PDB and making the PDBSite database[8]) is shown in the Figure 2. The ASPD amino acid frequency distribution shows the highest correlation with the codon number for each amino acid (the corresponding numbers for are: for ASPD – 0.83, for SwissProt – 0,73, for PDBSite - -0,14). This observation suggests a very important thing about native protein evolution – that the sequences of native proteins and their active sites are determined greatly by their evolutionary history and not by the functional requirements. It is illustrated by the fact that by means of phage display were retrieved small mimetics for large protein molecules (such as erythropoietin) [9, 10] that have no sequence similarity with them.

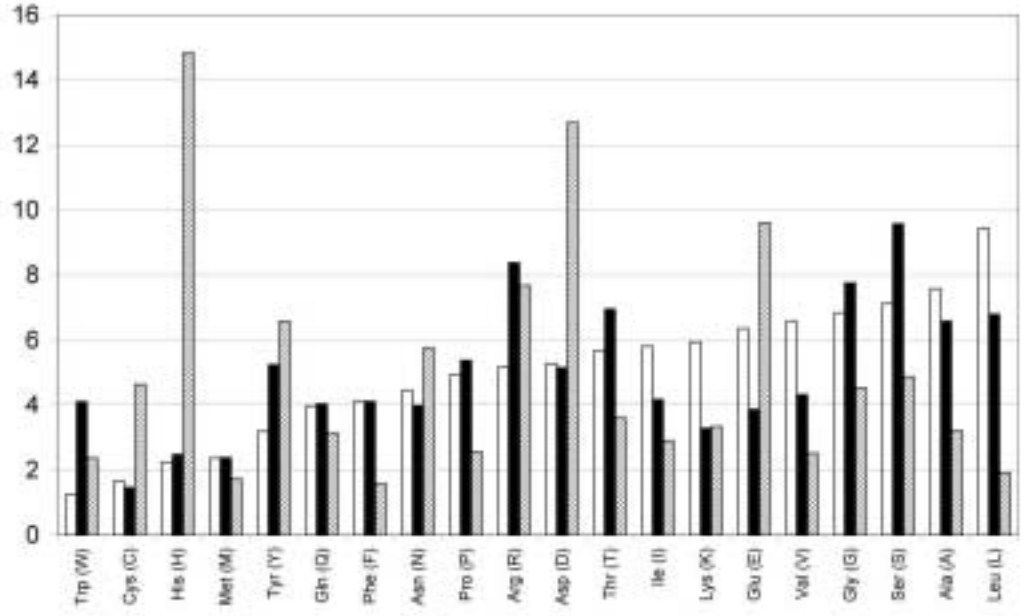


Figure 2 Amino acid composition of ASPD (black bars), SwissProt (empty bars), PDBSite (grey bars). Each amino acid content is given in percent, the aminoacids are ranged according to their content in SwissProt.

Amino acid similarity matrix.

We have calculated amino acid similarity matrix based on data stored in ASPD according to the procedure described in [11]. We have used the aligned sets of sequences to derive this matrix. In doing this, in each alignment we considered all the different sequences without weighting or clustering them and without regarding the number of times this sequence was found as an independent clone. We have then calculated the correlation coefficients of our matrix with other similar matrices and tested it in the BLAST homology search.

It should be noted, that the datasets on which our matrix and matrices of PAM [12] and BLOSUM [11] families were obtained are quite different by their nature. What is the set of proteins where we observe the interchanges of amino acids, is a set of proteins that perform exactly the same function under the same conditions. The set of proteins as they take it in Blocks database [13], on which BLOSUM matrices were derived, is a set of most conserved regions of distantly related proteins, that is to say the function they perform and the conditions of their existence are not exactly the same, and as diverge the sequences of those proteins, their exact function also can diverge. The PAM matrices, on the contrary, having been calculated on short evolutionary distances, consider mostly the mutable regions and may underestimate the probability of some amino acid substitutions. But in the light of differences in the initial dataset nature, it is not quite evident that the ASPD matrix (fig. 2) should correlate in any way with other matrices. Still our results show that it reaches the correlation of about 0.8 with matrices from BLOSUM family. The only one matrix which correlates with the ASPD matrix better is that of McLachlan, 1972 (correlation coefficient 0.82). The matrices from PAM family show the correlation of 0.71 – 0.73 with our matrix. The matrices

based on structural parameters generally yielded the correlation with the ASPD matrix in the range 0.4-0.8, and those based on hydrophobicity – even lower values (data is available at http://wwwmgs.bionet.nsc.ru/mgs/gnw/aspd/). One explanation for that may be that ASPD database (that is, phage display experiments) is biased in its content towards functional sites of proteins. Still there are a lot of experiments, where the positions crucial for structural integrity were randomized. The results of homology search for members of HIV envelope glycoprotein GP120 family (Pfam [14] identifier PF00516, 17891 members – the largest Pfam family, average length 128.5, average identity –10%) are shown in Table 2. These results show that our matrix is quite competitive when compared to the BLOSUM62 matrix, widely used in homology searches. So the general conclusion may be that in its overall character, evolution in vitro, manifested through amino acid interchanges, resembles that in vivo, despite the differences in amino acid frequencies, and there are no additional constraints which are present either only in vitro or in vivo, but that is not the case when only structurally important positions in native proteins are considered.

Table 2.

	BLOSUM62		ASPD matrix	
e-value	True positives	False positives	True positives	False positives
0.000001	16757	1	16951	1
0.00001	16832	1	16995	1
0.0001	16956	1	17027	8
0.001	17046	44	17066	41
0.01	17077	50	17096	51
0.1	17153	52	17146	56
1	17192	54	17191	65
10	17239	62	17257	109
100	17309	95	17332	354

References

1. Afonnikov D.A., Oshchepkov D.Yu., Kolchanov N.A. (2001) Detection of conserved physico-chemical characteristics of proteins by analysing clusters of positions with co-ordinated substitutions. *Bioinformatics,* to appear.
2. Chothia C. (1984) Principles that determine the structure of proteins. *Annu Rev Biochem*. **53**, 537-572.
3. Eisenberg D, Schwarz E, Komaromy M, Wall R. (1984) Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J Mol Biol*., **179**, 125-42.
4. Zimmerman J.M., Eliezer N., Simha R. (1968) The characterization of amino acid sequences in proteins by statistical methods. *J. Theor. Biol.,* **21**, 170-201.
5. Ponnuswamy PK, Prabhakaran M, Manavalan P. (1980) Hydrophobic packing and spatial arrangement of amino acid residues in globular proteins. *Biochim Biophys Acta*. **623,** 301-316.
6. Gu, H., Yi, Q., Bray,S.T., Riddle,D.S., Shiau,A.K. and Baker,D. (1995) A phage display system for studying the sequence determinants of protein folding *Protein Sci*., **4**, 1108-1117.
7. Kim,D.E., Gu,H. and Baker,D. (1998) The sequences of small proteins are not extensively optimized for rapid folding by natural selection *Proc. Natl. Acad. Sci. USA*, **95**, 4982-4986.
8. Ivanisenko D.A., Grigorovich D.A., Kolchanov N.A.// Proceedings of the Second International Conference on Bioinformatics of Genome Regulation and Structure/ Eds N. A. Kolchanov et al. Novosibirsk: ICG, 2000. V. 2. P. 171-174.

9. Wrighton,N.C., Farrell,F.X., Chang,R., Kashyap,A.K., Barbone,F.P., Mulchany,L.S., Johnson,D.L., Barrett,R.W., Jolliffe,L.K. and Dower,W.J. (1996) Small peptides as potent mimetics of the protein hormone erythropoietin *Science*, **273**, 458-463.
10. Wrighton,N. and Gearing,D. (1999) Fashioning FGFs from phage? *Nat. Biotechnol*., **17**, 1157-1158.
11. Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*, 89, 10915-10919.
12. Dayhoff, M. (1978) Atlas of protein sequence and structure. (Natl. Biomed. Res. Found., Washington), vol.5, suppl. 3, 345-358.
13. Jorja G. Henikoff JG, Greene EA, Pietrokovski S, and Henikoff S (2000) Increased coverage of protein families with the Blocks Database servers Nucl. Acids. Res., 28, 228-230.
14. Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, and Sonnhammer ELL (2000) The Pfam Protein Families Database Nucl. Acids. Res. 28, 263-266