Why inverse proteins are relatively abundant

Jean-Christophe Nebel* and Claude Godfrey Charles Walawage

Faculty of Computing, Information Systems and Mathematics, Kingston University, Kingston-upon-Thames, KT1 2EE, UK

*Address correspondence to this author at Faculty of Computing, Information Systems and Mathematics, Kingston University, Kingston-upon-Thames, KT1 2EE, UK; Tel: +44 20 8417 2740; Fax: +44 20 8417 2972; E-mail address: J.Nebel@kingston.ac.uk

Abstract: Studies have shown that inverse proteins are relatively abundant. In this work, we investigate the proposition that the repeat patterns they share with protein sequences explain this phenomenon. Using a new artificial set of peptide sequences which also display these features and a random set, we show that the presence of repeats contributes to protein sequence similarity. Further analysis confirms that most inverse proteins exhibit repeats. Therefore, we suggest the relative abundance of inverse proteins can be explained by the fact they display the same repeat structures and amino acid propensity of existing proteins.

Keywords: Inverse proteins, nullomers, random peptide chains, repeats.

INTRODUCTION

Driven by applications such as drug design, the extraction of biologically relevant features from protein sequences has become an essential task of bioinformatics. Methodologies have been developed to predict from sequence alone essential protein attributes including their secondary and 3D structures [1, 2], subcellular location [3,4], interaction partners [5], folding process [6] and functional annotation [7, 8]. Moreover, many web-servers are available providing user-friendly tools for in-silico analysis of proteins [9]. This paper investigates why inverse proteins are relatively abundant in order to gain valuable insight into protein sequence properties.

Many studies have focused on peptide chains created by inversing protein sequences[*]. These sequences are known as either 'inverse' or 'reverse' sequences [11, 12, 13]. Initially, it was assumed that, since amino acids share the same neighbours in a sequence and its inverse, they would fold into similar 3D structures. Early analysis of structural similarity discovered a few cases where the inverse protein had a mirrored protein structure [11]. However, a more comprehensive study established there is no correlation between the structures of a protein and its inverse [12]. Consequently, authors suggested inversion of protein sequences could be used to produce artificial sequences with similar amino acid propensity to real sequences, but folding like random proteins. On the other hand, inverse sequences are more common in nature than one would expect by chance [13]. Therefore, they must display some characteristics that are not present in random sequences. However, to date, the relative abundance of inverse proteins remains unexplained.

Another line of research has focused on sequence repeats [14]. Andrade et al. have highlighted the importance of duplication in protein evolution [15]. In particular, the study of the evolution of multi-domain proteins revealed that around 30% of them evolved through repetitions [16]. Moreover, it was shown around 35% of long structural repeats display a 2-fold symmetry that corresponds to a homodimer configuration [17]. Furthermore, experimental work discovered that creation of repetitions within a random sequence tends to produce more proteins with secondary structures than random sequences [18]. This led to the suggestion that repetition could be a process allowing the generation of De novo proteins.

In this work, we investigate the proposition that inverse peptide chains are more common than random ones mainly because they display periodicity and repeat patterns present in protein sequences. First, we have created a new artificial reference set of peptide sequences – 'opproteins' - which also shares these features without preservation of either amino acid composition or neighbourhood as in the case of inverse peptide chains. Then, after querying protein sequence database for similarity, we compared results with those obtained with inverse and random sets.

MATERIALS AND METHODS

New artificial peptide sequences: opproteins

In order to test the hypothesis that the abundance of inverse proteins can be explained by the fact that inversion preserves repetition properties of a protein sequence, we needed to design a new artificial peptide dataset sharing this feature. This can be achieved by taking a representative sample of known proteins and applying a global operation where each given amino acid is replaced by a different one. However, since substitution of a residue by another tends to be neutral if they share similar properties, a scheme where replacements were random could lead to the creations of peptides which could still be aligned with their proteins of references. To prevent this and prove that repetition instead of amino acid environment is the main explanation behind the abundance of inverse proteins, we propose to generate a set of peptide sequences sharing repeat patterns with reference proteins, but displaying very different amino acid environments.

We introduce the concept of 'opprotein': an 'opprotein' is the peptide chain which is the most unlike a given protein sequence, i.e. its opposite. More specifically, the opposite of a protein sequence, P, or 'opprotein', $\ominus P$, is defined as the amino acid sequence where each residue of P is replaced by an amino acid with the most opposite physico-chemical properties as defined by a given substitution matrix. However, since a few amino acids such as tryptophan, proline and aspartic acid are the most opposite to several others, a simple substitution of each residue by the most different would not produce a sequence composed of 20 different amino acids. In order to ensure that each residue is replaced by a different one, an optimisation algorithm was used to produce the optimal substitution table in term of replacing the 20 amino acids by their opposites.

[*] Inverse sequences should not be confused with 'anti sense' proteins where the inverse of coding DNA sequences are used to produce an inverse peptide chain of complementary codons [10].

Generation of opproteins

The optimisation task that needs to be performed is defined as producing the optimal opposite list of pairs of amino acids using a given substitution matrix as a cost function quantifying the oppositeness between all pairs of residues. In this work, the popular BLOSUM62 matrix was used [19], e.g. it is the default matrix in BLAST [20], but any substitution matrix would be equally suitable.

This cost minimization problem corresponds to a typical assignment problem, which was solved by Kuhn in 1955 using the Hungarian algorithm [21]. It is important to note that, in term of oppositeness cost, a substitution matrix is not symmetrical: although the substitution cost between A and B equal the cost between B and A, the most opposite residue of A may be B, whereas the most opposite residue of B could be C. For example, according to BLOSUM62, tryptophan is the amino acid which is the most different from serine with a value of -3, while asparagines, proline and aspartic acid are the most different from tryptophan with a value of -4. In this case, serine is only ranked 4th in tryptophan's list of most different amino acids.

Table 1 shows the amino acid substitution table generated from BLOSUM62 using the Hungarian algorithm. For each substitution the ranking of the substitute is also provided. 12 substitutions are symmetric and the average ranking of the chosen amino acid in the list of most different amino acids is 2.05.

Table **1**. Amino acid substitution table generated from BLOSUM62 - symmetric substitutions are shown in bold.

Residue	**A**	**R**	**N**	D	**C**	**Q**	**E**	G	**H**	**I**	L	K	M	F	P	**S**	**T**	**W**	Y	**V**
Opposite residue	**N**	**V**	**A**	M	**E**	**I**	**C**	L	**T**	**Q**	D	F	G	P	Y	**W**	**H**	**S**	K	**R**
Rank in BLOSUM62	2	1	6	3	1	1	1	1	5	2	1	1	1	1	3	1	1	4	4	1

Since all substitutions are not symmetric, the opposite of the opprotein of P, $\ominus P$, is not P: $\ominus\ominus P \neq P$. However, as illustrated in the following alignments, $\ominus\ominus\ominus\ominus P = P$, and $\ominus\ominus P$ and P are very similar. Experiment using PAM250 [22] instead of BLOSUM62 also shows that $\ominus\ominus\ominus\ominus P = P$.

```
P           ARNDCQEGHILKMFPSTWYV
θθP         ARNGCQEDHIMPLYKSTWFV
θθθθP       ARNDCQEGHILKMFPSTWYV
            ***.***.**: :: ***:*
```

Datasets

Datasets used for this study consist of a set of representative protein sequences, its conversions to opproteins and inverse sequences, and a set of random peptide sequences. Representative protein sequences were extracted from the PDB [23] on 16th July 2009 using the advanced search option of RCSB PDB [24] selecting entries containing only 1-chain proteins of at least 100 amino acids and trimmed so that no single pair of proteins has sequence identity higher than 30%. 5489 sequences were returned with an average length of 263 residues. We call this reference set PDB30.

In order to produce opproteins and inverse sequences, we created a Java applet [25] which reads a file containing protein sequences and converts them into either inverse proteins or opproteins using the Hungarian algorithm [21] and a specified substitution matrix. The set of opproteins, $\ominus$PDB30, was generated using PDB30 as the input file and BLOSUM62 [19]. Similarly, the set of inverse sequences, *inv*PDB30, was created by inversing all sequences of PDB30.

Finally, a set of random peptide chains, *ran*5489, was generated using a random sequence generator which allocates the same propensity to all residues [26]. This set is composed of 5489 sequences of length 263 residues so that it can be compared with the other artificial sets derived from PDB30.

Figure 1 shows the average amino acid propensities found in the three artificial datasets and PDB30 which is used as reference (by definition PDB30 and *inv*PDB30 have identical amino acid propensities). The figure reveals that the residue composition of opproteins is particularly unnature-like.

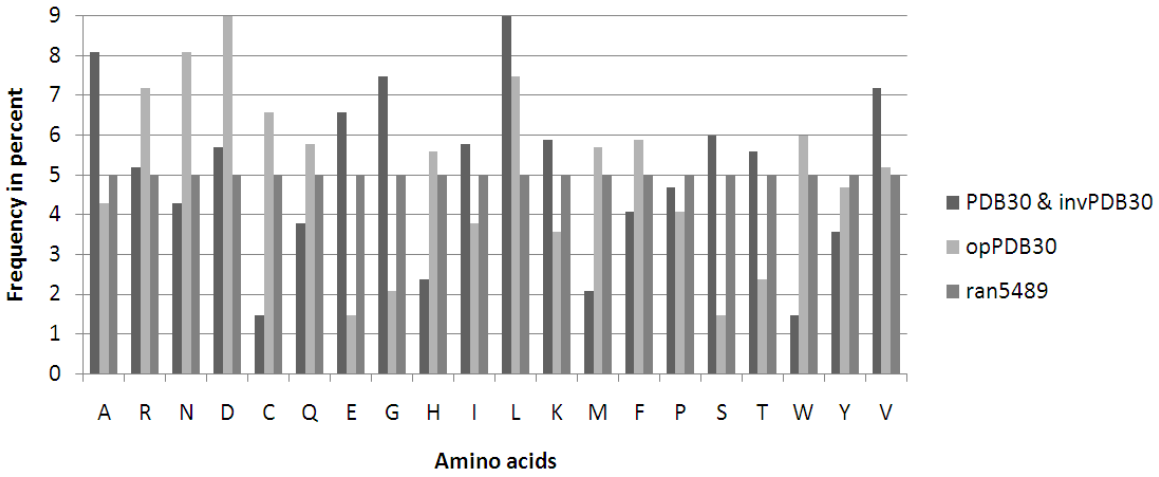


Figure **1**. Amino acid propensities found in PDB30 & *inv*PDB30, ⊖PDB30 (*op*PDB30) and *ran*5489.

Methodology

Sequences of the three artificial datasets, i.e. *inv*PDB30, ⊖PDB30 and *ran*5489, were analysed to establish if they show similarity to existing proteins. Each sequence was processed by BLASTP 2.2.20 using standard parameters against their non-redundant protein database downloaded on 17th July 2009 (9,298,190 entries) [27]. Then, among the retrieved protein sequences the lowest E-value, if any, was extracted.

RESULTS AND DISCUSSION

Amino acid environment of opproteins

In order to illustrate that the amino acid environment of opproteins in ⊖PDB30 can be significantly different from those found in PDB30, *inv*PDB30 and *ran*5489, we looked for the presence of amino acid sequences which have never been found in any protein chain (including inversed chains), or 'nullomers' [28]. It has been hypothesised they are signatures of natural selection against deleterious sequences [28, 29], but their significance is still under investigation [30]. According to Hampikian et al.'s web site [31], as in October 2009, 38 absent length-5 amino acid sequences have been identified as nullomers [28]. Assuming that each sequence of 5 residues is equiprobable, the number of nullomers which are expected to be found in a random dataset of 5489 peptide sequences of length 263 can be calculated using the binomial distribution to represent their discrete probability distribution. In a random dataset, the expected number of hits is 16.9 with a standard deviation, σ, of 4.1.

Scanning of *ran*5489 reveals the presence of 16 nullomers, which confirms its randomness. 37 nullomers were found in ⊖PDB30. This demonstrates that opproteins display amino acid environments which cannot be found in nature. Moreover, since a distance of 4.9 σ from the expected value is highly significant, this proves that opproteins are not random proteins.

Abundance of inverse proteins

Following the processing of the artificial peptide sets by BLASTP, we extracted the E-value of the first hit for each sequence. Figure 2 is a plot of these first hit E-values after clustering them into different bins: each bin accumulates values between consecutive powers of 10, i.e. bin -*i* contains E-values ranging from $1x10^{-i-1}$ to $1x10^{-1}$ excluded. The last bin, called '1+', contains all sequences which did not receive any hit with an E-value smaller than $1x10^{1}$. Consequently, for each test set, the sum of all hits recorded on the graph is equal to 5489, i.e. the number of sequences per set. In datasets of size 5489, according to standard E-value calculations [32], one would theoretically expect in our experiments 5 and 50 random hits, respectively, in bin '-3' and '-2'. In practice, our results are more conservative since the first 14 hits of the random set appear in bin '-2'. Consequently, in our analysis we will consider that hits with an E-value$<1x10^{-3}$ are significant.

With 19 significant hits in *inv*PDB30, Figure 2 confirms results of previous studies [13]: inverse sequences are much more protein-like than random sequences. When E-value$>1x10^{-3}$, the opprotein hit profile is generally close to the random sequence one. However, Figure 2 also reveals that, in 5 instances, there is a very high similarity between opproteins of ⊖PDB30 and a real protein. Remarkably, 4 protein sequences produced both inverse and opproteins in this list, Table 2. In view of these results, one may suggest that opproteins share some properties with inverse and real sequences, which are not found in random sequences.

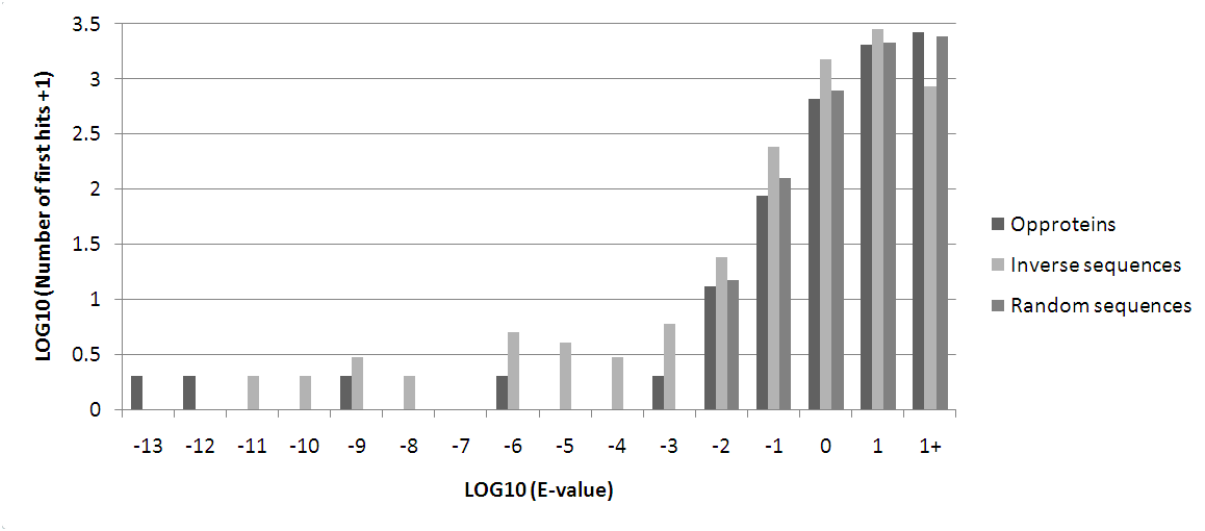


Figure **2**. Similarity of artificial peptide sequences with real protein sequences.

Initial investigation of significant hits revealed that 3 inverse sequences match the chain 1C94|A, which was artificially engineered to study the inverse structure of gcn4 leucine zipper [33]. Therefore, they were excluded from further analysis. The remaining sequences were then processed by Swelfe v1.0 [17] using default parameters to detect significant internal repeats (p-value < 0.01). Information related to the presence of repeats was also extracted from publications associated with the proteins of interest. These data are shown in Table 2.

Table **2**. List of proteins whose either inverse or opposite sequence has shown some significant similarity to a protein sequence, associated E-values and overlaps with the first hit and numbers of significant hits, Swelfe scores and repeat related features.

PDB code	Length	**Inverse**	Overlap	Hits	**Opprotein**	Overlap	Hits	**Swelfe score**	**Repeat related feature**
3DU1:X	257	5E-12	82.5%	68	6E-14	82.1%	17	145	Pentapeptide repeat protein
2J8K:A	201	1E-10	82.6%	74	1E-13	73.6%	43	133	Fusion of 2 pentapeptide repeat proteins
3G06:A	622	1E-11	30.7%	6	5E-10	25.7%	5	257	Leucine-rich repeat protein
3BZ5:A	457	2E-10	54.0%	2				218	15 leucine-rich repeats
1OZN:A	285	3E-09	92.3%	1				104	Leucine-rich repeat module
2GY5:A	423	2E-07	40.7%	22				86	3 immunoglobulin domains
3CU9:A	314	6E-07	40.1%	1				0	Five-bladed-beta-propeller fold domain
2O6W:A	150	1E-06	62.7%	5	8E-07	56.7%	5	74	Pentapeptide repeat protein
1N7D:A	699	2E-06	45.8%	1				132	Two beta propeller modules
2JF2:A	264	1E-05	69.3%	1				72	10 left-handed beta-helix coils, each composed of 3 hexapeptide repeats
2Q7Z:A	1931	1E-05	20.9%	7				1800	30 short complement regulator domains
1ZIW:A	680	2E-04	60.9%	1				156	23 leucine-rich repeats
7ODC:A	424	2E-04	35.8%	1				0	Alpha/beta barrel domain
3GAU:A	1213	4E-04	27.2%	1				324	20 short complement regulator domains
2I1J:A	575	5E-04	29.6%	1				0	/
2G5D:A	422	7E-04	30.3%	1				0	Double psi beta-barrel
2O2G:A	223				8E-04	79.8%	1	0	/

Most of these proteins, i.e. 12 out of 17, display significant repeats as defined by Swelfe [17]. Moreover, proteins without repeats tend to produce either inverse sequences or opproteins with lower E-values and only receive one single significant hit. Literature information also confirms the presence of repeats or repeating structural elements, i.e. beta-barrel, beta propeller and duplicated folds, in most cases.

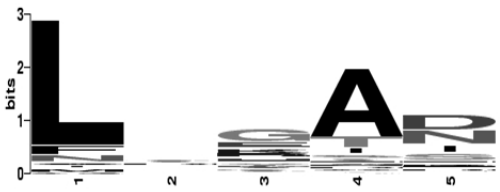
In order to illustrate the effect of repeats on inverse and opprotein sequences, we present a detailed analysis of HetL (PDB ID code 3DU1:X), a protein involved in regulation of heterocyst differentiation [34]. Figure 3 a) and c) highlight, respectively, the pentapeptide repeat which composes its sequence and the associated structural repetitions. Multiple alignment of all these pentapeptides allows the generation of a well conserved pattern as represented by its consensus logo [35], Figure 3 b).

a)

>3DU1:X

---HM	*NVGEI*	*LRHYA*	AGKRN	FQHIN	**L**QEIE	**L**TNAS	**L**TGAD	**L**SYAD	**L**RQTR
LGKSN	**F**SHTC	**L**READ	**L**SEAI	**L**WGID	**L**SEAD	**L**YRAI	**L**READ	**L**TGAK	**L**VKTR
LEEAN	**L**IKAS	**L**CGAN	**L**NSAN	**L**SRCL	**L**FQAD	LRP**S**S	NQRTD	**L**GYVL	**L**TGAD
LSYAD	**L**RAAS	**L**HHAN	**L**DGAK	**L**CRAN	FGR**T**I	**Q**WG-N	LA-AD	**L**SGAS	**L**QGAD
LSYAN	**L**ESAI	**L**RKAN	**L**QGAD	**L**TGAI	**L**KDAE	**L**KGAI	MPDGS	IH	

b)



c)

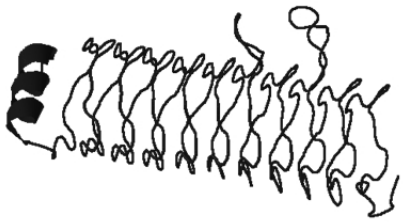


Figure **3**. Periodicity of the 3DU1:X protein: a) sequence (amino acids belonging to the initial alpha helix are in italic, those starting a beta turn are in bold), b) consensus logo and c) 3D structure

Since inverse and opprotein sequences conserve repetition patterns, in this instance, they can also be described by a 5-residue long consensus logo, Figure 4 a) and b) respectively. Figure 4 c), e), g), i) and k) display the 5-residue long descriptors associated with the first 5 significant hits of the inverse sequence of 3DU1:X. The logos reveal that the first 3 hits contain extremely well conserved pentapeptides forming more than 50% of these protein sequences. The 2 other hits involve proteins with a repeat pattern of 10 residues, which only covers 10% of their long sequences. Individual alignments of those consensus logos with the descriptor of the inverse of 3DU1:X demonstrate why this inverse sequence obtains hits with such low E-values: up to 4 of the consensus positions describing the hit proteins match the logo of the inverse sequence.
Similarly, Figure 4 d), f), h), j) and l) present the consensus logos of the proteins which are the most similar to the opprotein of 3DU1:X. They are all made of well conserved pentapeptides where the first residue is most likely to be an aspartic acid, which is in consistent with the descriptor of the opprotein. It is interesting to note that despite their very different origins, both the inverse and the opprotein sequences are matched by the same protein (XP_002169517.1).

Discussion

In agreement with some experimental work [18] which shows the presence of repeating elements within a peptide sequence tends to make the amino acid chain more protein-like, our study suggests that repeats are the main contributor of the abundance of inverse proteins. However, since inverse proteins are more common than opproteins that share the same repeat patterns, other factors must be involved. Unlike opproteins, inverse peptide sequences have the same residue propensity as known proteins. Thus, inverse sequences are intrinsically more similar to real proteins than opproteins. Consequently, amino acid distribution must also play a part in the similarity between a peptide chain and a protein. Experiments conducted in this study cannot determine if local residue environments, which are arguably conserved in inverse sequences, also have an impact on a sequence to be protein-like. This issue could be addressed in future work, for example, by comparing results obtained by random sequences produced from an amino acid distribution against random sequences generated piecewise from short amino acid sequences extracted from real proteins.

In conclusion, we propose the relative abundance of inverse proteins can mainly be explained by the fact they display the same repeat structures and amino acid propensity of existing proteins. A consequence is that the use of inverse sequences as a negative set in experiments should be done with caution as they cannot be considered as random.

ACKNOWLEDGEMENTS

The authors would like to thank Witold Dyrka for helpful discussions.

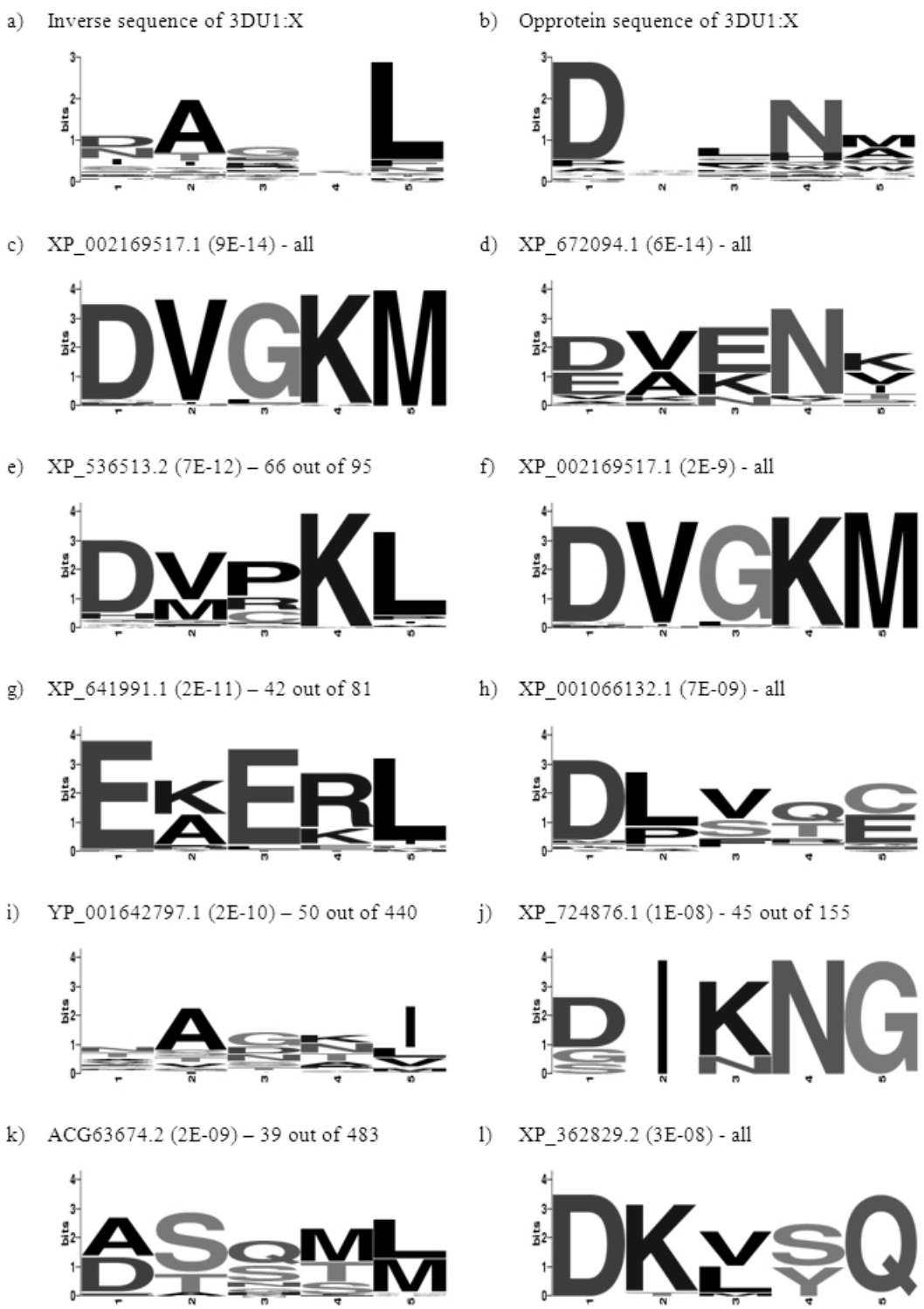


Figure **4**. Consensus logos describing the a) inverse and b) opprotein sequences of 3DU1:X and their 5 best hits c), e), g), i), k) and d), f), h), j), l) respectively. In addition to their NCBI Reference id, each hit is provided with its E-value and the number of pentapeptides which was used to create the logo.

REFERENCES

[1] Konopka, B.; Dyrka, W.; Nebel, J.-C.; Kotulska, M. Accuracy in predicting secondary structure of ionic channels. In: *New Challenges in Computational Collective Intelligence*, Studies in Computational Intelligence, Ed.; Springer-Verlag, **2009**, 315-326.

[2] Zhanga, Y. Protein structure prediction: when is it useful? *Current Opinion in Structural Biology*, **2009**, *19*(2), 145-155.

[3] Lin, H.; Chen, C.; Sung, T.; Ho, S.; Hsu, W. Protein subcellular localization prediction of eukaryotes using a knowledge-based approach. *BMC Bioinformatics*, **2009**, *10*(15), S8.

[4] Chou, K. C.; Shen, H. B. Review: Recent progresses in protein subcellular location prediction. *Anal. Biochem.*, **2007**, *370*, 1-16.

[5] Shoemaker, B.; Panchenko, A. Deciphering protein–protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PLoS Comput. Biol.*, **2007**, *3*, e43.

[6] Chou, K. C.; Shen, H. B. FoldRate: A web-server for predicting protein folding rates from primary sequence, *The Open Bioinformatics Journal*, **2009**, *3*, 31-50.

[7] Juncker, A.; Jensen, L.; Pierleoni, A.; Bernsel, A.; Tress, M.; Bork, P.; von Heijne, G.; Valencia, A.; Ouzounis, C.; Casadio, R.; Brunak, S. Sequence-based feature prediction and annotation of proteins. *Genome Biol.*, **2009**, *10*(2),206.

[8] Dyrka, W.; Nebel, J.-C. A Stochastic Context Free Grammar based Framework for Analysis of Protein Sequences, *BMC Bioinformatics*, **2009**, *10*, 323.

[9] Chou, K. C.; Shen, H. B. Review: recent advances in developing web-servers for predicting protein attributes. *Natural Science*, **2009**, *2*, 63-92.

[10] Chou, K. C.; Zhang, C. T.; Elrod, D. W. Do antisense proteins exist? *J. Protein Chem.*, **1996**, *15*, 59-61.

[11] Guptasarma, P. Reversal of peptide backbone direction may result in the mirroring of protein structure. *FEBS Lett.*, **1992**, *310*, 205-210.

[12] Lorenzen, S.; Gille, C.; Preißner, R.; Frömmel, C. Inverse sequence similarity of proteins does not imply structural similarity. *FEBS Lett.*, **2003**, *545*(2-3), 105-9.

[13] Preißner, R.; Goede, A.; Michalski, E.; Frömmel, C. Inverse sequence similarity in proteins and its relation to the three-dimensional fold. *FEBS Lett.*, **1997**, *414*(2), 425-429.

[14] Chou, K. C.; Heinrikson, R. L. Prediction of the tertiary structure of the complement control protein module. *J. Protein Chem.*, **1997**, *16*, 765-773.

[15] Andrade, M.; Perez-Iratxeta, C.; Ponting, C. Protein repeats: Structures, functions, and evolution. *J. Struct. Biol.*, **2001**, *134*, 117–131.

[16] Björklund, A.K.; Ekman, D.; Light, S.; Frey-Skött, J.; Elofsson, A. Domain Rearrangements in Protein Evolution. *J. Mol. Biol.*, **2005**, *353*(4), 911-23.

[17] Abraham, A.-L.; Rocha, E.; Pothier J. Swelfe: a detector of internal repeats in sequences and structures. *Bioinformatics*, **2008**, *24*(13), 1536-1537.

[18] Shiba, K.; Takahashi, Y.; Noda, T. On the role of periodism in the origin of proteins. *J. Mol. Biol.*, **2002**, *320*, 833-840.

[19] Henikoff, S.; Henikoff, J.G. Amino Acid Substitution Matrices from Protein Blocks. *PNAS*, **1992**, *89*, 10915–10919.

[20] Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.*, **1990**, *215*, 403-410.

[21] Harold, W.K. The Hungarian Method for the assignment problem. *Naval Research Logistics Quarterly*, **1955**, *2*, 83–97.

[22] Dayhoff, M.O.; Schwartz, R.; Orcutt, B.C. A model of Evolutionary Change in Proteins, *Atlas of protein sequence and structure*, **1978**, *5*(3), 345–358.

[23] Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The Protein Data Bank. *Nucleic Acids Research*, **2000**, *28*, 235-242.

[24] Berman, H.M.; Henrick, K.; Nakamura, H. Announcing the worldwide Protein Data Bank. *Nature Structural Biology* , **2003**, *10*(12), 980.

[25] Nebel, J.-C.; Walawage, C.G. Protein Sequence Converter http://staffnet.kingston.ac.uk/~ku33185/ProteinSequenceConverter/OpProtein.html (Accessed February 9, **2010**).

[26] Stothard, P. The Sequence Manipulation Suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *Biotechniques*, **2000**, *28*, 1102-1104.

[27] Altschul, S.F.; Madden, T.L.; Schaffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **1997**, *25*, 3389-3402.

[28] Hampikian, G.; Andersen, T. Absent sequences: Nullomers and primes. In: *Pacific Symposium on Biocomputing*, **2007**, *12*, 355-366.
[29] Tuller, T.; Chor, B.; Nelson, N. Forbidden penta-peptides, *Protein Sci.*, **2007**, *16*(10), 2251–2259.
[30] Acquisti, C.; Poste, G.; Curtiss, D.; Kumar, S. Nullomers: really a matter of natural selection? *PLoS ONE*, **2007**, *2*(10).
[31] Boise State University. Nullomer Project: SeqCount. http://trac.boisestate.edu/bioinformatics/nullomers/ (Accessed December 1, **2009**).
[32] Altschul, S.; Gish, W. Local alignment statistics. *Meth. Enzymol.*, **1996**, *266*, 460-480.
[33] Mittl, P.R.; Deillon, C.; Sargent, D.; Liu, N.; Klauser, S.; Thomas, R.M.; Gutte, B.; Grütter, M.G. The retro-GCN4 leucine zipper sequence forms a stable three-dimensional structure. *Proc. Natl. Acad. Sci.*, **2000**, *97*(6), 2562-6.
[34] Ni, S.; Sheldrick, G.; Benning, M.; Kennedy, M. The 2A resolution crystal structure of HetL, a pentapeptide repeat protein involved in regulation of heterocyst differentiation in the cyanobacterium Nostoc sp. strain PCC 7120. *J. Struct. Biol.*, **2009**, *165*, 47-52.
[35] Crooks, G.; Hon, G.; Chandonia, J.; Brenner, S. WebLogo: A sequence logo generator, *Genome Research*, **2004**, *14*, 1188-1190.