Large cryptic internal sequence repeats in protein structures from *Homo sapiens*

R Sarani[1], N A Udayaprakash[1], R Subashini[1], P Mridula[1], T Yamane[2] and K Sekar[1,3,*]

[1]*Bioinformatics Centre, Indian Institute of Science, Bangalore 560 012, India*
[2]*Department of Biotechnology, Graduate School of Engineering, Nagoya University, Furo-cho, Nagoya 464-8603, Japan*
[3]*Supercomputer Education and Research Centre, Indian Institute of Science, Bangalore 560 012, India*

**Corresponding author (Fax, +91-80-23600683/23600551; Email, sekar@serc.iisc.ernet.in, sekar@physics.iisc.ernet.in)*

Amino acid sequences are known to constantly mutate and diverge unless there is a limiting condition that makes such a change deleterious. However, closer examination of the sequence and structure reveals that a few large, cryptic repeats are nevertheless sequentially conserved. This leads to the question of why only certain repeats are conserved at the sequence level. It would be interesting to find out if these sequences maintain their conservation at the three-dimensional structure level. They can play an active role in protein and nucleotide stability, thus not only ensuring proper functioning but also potentiating malfunction and disease. Therefore, insights into any aspect of the repeats – be it structure, function or evolution – would prove to be of some importance. This study aims to address the relationship between protein sequence and its three-dimensional structure, by examining if large cryptic sequence repeats have the same structure.

[Sarani R, Udayaprakash N A, Subashini R, Mridula P, Yamane T and Sekar K 2009 Large cryptic internal sequence repeats in protein structures from *Homo sapiens*; *J. Biosci.* **34** 103–112]

1. Introduction

A repeat is defined as two or more contiguous segments of amino acid (three or more) residues with identical and similar sequence. When such repeats are in high-complexity regions, they are called 'cryptic' (Tantz *et al* 1986). Although low-complexity repeats are essential for evolutionary analysis and comprise a large section of the eukaryotic genome, high-complexity repeats are usually associated with a particular structure or function. This study considers large cryptic repeats comprising eight or more residues, as Hancock *et al* (2001) fixed the length of a moderate-sized repeat as being between five and eight amino acids. The study of repeats is crucial because all but 5–6% of the high eukaryotic genome is repetitive (Hancock and Simon 2005). Internal protein repeats are observed to be associated with structural motifs or domains. It is evolutionarily more 'economical' to evolve complex structures such as multiple domains by using 'modular plug-ins' (Heringa 1998) to fulfil a specific function. Furthermore, longer repeats normally act to enhance the stability of the native fold of the protein and, while small repeats interact with each other, larger repeats may either interact or remain isolated like beads on a string (Heringa 1998).

Three prominent reviews on repeats are those of Heringa (1998), Marcotte *et al* (1999) and Andrade *et al* (2001), and they concentrate on the relationship between structural repeats and their primary structure along with

Keywords. Propensity; structure–function correlation; human genome; structural plasticity; three-dimensional structure; identical and similar sequence repeats

Abbreviations used: ADP, adenosine diphosphate; ATP, adenosine triphosphate; 3dSS, 3-dimensional structural superposition; FAIR, finding all internal repeats; NCBI, National Centre of Biotechnology Information; PAM, Point Accepted Mutations; PDB, Protein Data Bank; PSAP, Protein Structure Analysis Package; RMSD, root mean square deviation

the characteristics of protein families. Andrade *et al* (2001) discuss the evolution of repeats as modules in the proteins. It is mentioned that the number of repeats in a protein can vary between orthologous proteins, implying that the loss or gain of repeats is very rapid in evolution. In fact, there exists a class of structural repeats which can occur in non-integer multiples and whose boundaries do not coincide with the secondary structural elements, such as the β-propellers (Smith *et al* 1999). However, once the number of repeats has been established, sequence similarity between the repeats tends to decrease as the repeats begin to diverge rapidly. Thus, it is possible that functional constraints exist on the assembly rather than on the individual repeats. A study on the role of short polyamino repeats (Djian 1998) brought out novel ideas on the role of repeats in neurodegenerative disease. Another group studied simple sequence repeats of less than six residues and their implications in network evolution (Hancock and Simon 2005). These studies confirmed and contradicted several of the points raised in an earlier work (Andrade and Bork 1995). However, none of these studies examines the role of conserved sequence repeats. Hence, there is a need to better understand the sequence, structure and function of repeats.

In this study, two kinds of repeats are considered. In 'identical repeats', which arise due to duplication of DNA, no mismatch is allowed. Any amino acid sequence could undergo mutation and thus repeats are expected to diverge. Conservation of repeats would be found in cases where either the repeat is newly duplicated or it has a structural and/or functional purpose. Larger repeats are less likely to be conserved without a specific purpose. On the other hand, 'similar repeats' allow specific mismatches, i.e. the substitution of structurally similar amino acids (for example: F ↔ Y; S T; V T; L I; K R; D N; Q E). One example of a similar repeat is from vasodilator-stimulated phosphoprotein (from PDB-id: 1USE), where '|' represents an identical match and ':' is a mismatch between structurally similar amino acids, as described above.

```
LQRVKQELLE
|  |: | | : |: : |
LQKVKEEIIE
```

2. Materials and methods

The human genome sequence was downloaded from the National Centre of Biotechnology Information (NCBI) ftp site. To identify the corresponding three-dimensional protein structures of the human genome available in the Protein Data Bank (PDB), every sequence of the NCBI dataset was used as a query sequence against all the protein sequences available in the PDB using PSI-BLAST (Altschul *et al* 1997). A 90% sequence cut-off was used. Using this procedure, a total of 3136 non-redundant structures from *Homo sapiens* was obtained, which comprised 5796 protein chains. This study makes extensive (and exclusive) use of the algorithm FAIR (finding all internal repeats; Banerjee *et al* 2008), to find internal sequence repeats. FAIR was developed based on simple dynamic programming concepts. In order to find internal repeats within a sequence, it aligns the sequence on the X and Y axes. Next, it finds the suboptimal alignments and, finally, it displays the repeat along with the location after weeding out repeats that are merely subsets of larger repeats. After the repeats were found, a web server, three-dimensional structural superposition (3dss) (Sumathi *et al* 2006) was used to superimpose the three-dimensional structures of the repeats and obtain the structural alignment. Information about the protein was obtained from the Protein Structure Analysis Package (PSAP; Balamurugan *et al* 2007) and necessary three-dimensional atomic coordinates for the protein molecules used in the present study were obtained from the anonymous FTP server maintained at the Bioinformatics Centre, Indian Institute of Science, Bangalore, India. Further calculations and necessary analyses were carried out using locally developed Perl scripts.

3. Results and discussion

Cryptic repeats comprising eight or more amino acid residues are included in this study. Out of the entire dataset, only 19 proteins were found to have 38 identical sequence repeats (table 1), while 30 proteins had 45 similar repeats (table 2) of eight or more residues each. Interestingly, almost all the identical and similar repeats were found to have the same three-dimensional structure. Out of the 38 large cryptic identical repeats found (table 1), only two did not superimpose (from PDB-ids 1JBQ and 1FYH) since the atomic coordinates are missing in their PDB file. For the same reason, two repeats from the 45 similar repeats found (table 2) did not superimpose (from PDB-id 2NQ3 and 1FYH). In fact, it is intriguing that although large-module repeats ought to exist in the proteins, none apart from the one in interferon-γ (PDB-id 1FYH) have remained so highly conserved with respect to the sequence. We also found that, for large repeats, flanking residues (excluding Glycine and Proline) did not influence the three-dimensional structure. It is likely that identical and similar repeats serve some useful biological function, such as activity or scaffolding. This is supported by the fact that the amino acid sequence is highly conserved only in the case of some exacting function of the structure of the protein. An example of this can be seen in ice-binding β-sheets of insect anti-freeze protein (Liou *et al* 2000). Proteins with repeats conserved across species are under strong purifying selection (Hancock *et al* 2001). Thus, large conserved repeats have properties of selectively conserved rather than neutral sequences.

Table 1. Large identical repeats from the non-redundant dataset of *Homo sapiens* proteins

PDB-id and chain	Name of protein	Repeat length[$]	Repeat	Location	Location	STAMP score	RMSD (Å)
1CZA N	Hexokinase I	10 (E)	GFTFSFPCQQ	151–160	599–608	9.781	0.179
		14 (E/H)	VAVVNDTVGTMMTC	204–217	652–665	9.755	0.266
1E07 A	Carcinoembryonic antigen	10	VILNVLYGPD	195–204	373–382	9.777	0.189
		10	QNTTYLWWVN	316–325	494–503	9.757	0.237
		13	QSLPVSPRLQLSN	327–339	505–517	9.791	0.090
		15	LSCHAASNPPAQYSW	223–237	401–415	9.723	0.254
		21	TYLWWVNNQSLPVSPRLQLSN	141–161	319–339	9.799	0.030
		27	SWLPVSPRLQLSNGNNRTLTLFNVT RND	149–175	505–531	9.795	0.070
		28	ELPKPSISSNNSKPVEDKDAVAFTC EPE	109–136	465–492	9.048	0.242
1JBQ A	Cystathionine β-synthase	14 (E)	GIPSETPQAEVGPT	22–35	22–35	–	–
1KI0 A	Angiostatin	8 (E)	ENYCRNPD	50–57	222–229	9.774	0.180
1L6J A	Matrix metalloproteinase	13 (T)	SYSACTTDGRSDG	221–233	279–291	9.675	0.454
1FYH A	Interferon-γ	9 (H)	ELIQVMAEL	113–121	237–245	9.683	0.366
		106 (H)	VKEAENLKKYFNAGHSDVADNGTL FLGILKNWKEESDRKIMQSQIVSFYF KLFKNFKDDQSIQKSVETIKEDMNV KFFNSNKKKRDDFEKLTNYSVTDLN VQRKAI	6–111	130–225	–	–
1LAR A	LAR protein	8 (H)	MVQTEDQY	258–265	549–556	9.792	0.095
		9 (E)	AYIATQGPL	87–95	376–384	9.717	0.362
		11(E/H)	VHCSAGVGRTG	214–224	505–515	9.735	0.289
1M9I A	Annexin VI	8 (bend)	KAMKGLGT	261–268	373–380	9.629	0.544
		8 (H)	RIMVSRSE	275–282	623–630	9.769	0.202
1N11 A	Ankyrin	8 (H)	GLTPLHVA	405–412	569–576	9.785	0.147
1OI1 A	SCML2 protein	9 (T)	FKVGMKLEA	41–49	150–158	9.769	0.211
2CMR A	Transmembrane glycoprotein	42(H)	QLLSGIVQQQNNLLRAIEAQQHLLQL TVWGIKQLQARILAGG	2–43	178–219	9.140	0.573
		130 (H)	QLLSGIVQQQNNLLRAIEAQQHLLQ LTVWGIKQLQARILAGGSGGHTTW MEWDREINNYTSLIHSLIEESQNQQE KNEQELLEGSSGGQLLSGIVQQQNN LLRAIEAQQHLLQLTVWGIKQLQAR ILAGG	2–131	90–219	9.140	0.573
1OZ2 A	Lethal brain tumour-like protein	8 (E)	MKLEAVDR	153–160	257–264	9.759	0.282
1OZN A	Reticulon 4 receptor	8 (E)	FLHGNRIS	38–45	159–166	9.786	0.097
2EW9 A	Cu-transporting ATPase2	8 (E/H)	GMTCASCV	12–19	88–95	8.106	1.539
2NZT A	Hexokinase II	9 (H)	FEKMISGMY	279–287	729–735	9.782	0.160
		9 (bend)	NMEWGAFGD	244–252	692–700	9.785	0.138
		10 (H/E)	SEDGSGKGAA	431–440	879–888	9.778	0.147
		11 (H)	LGFTFSFPC	134–144	582–592	9.735	0.317

Table 1. (*Continued*)

PDB-id and chain	Name of protein	Repeat length[$]	Repeat	Location	Location	STAMP score	RMSD (Å)
		14 (E)	FLALDLGGTNFRVL	66–79	514–527	9.770	0.197
		15 (H/E)	VAVVNDTVGTMMTCG	190–204	638–652	9.645	0.521
		17(H/E)	GLIVGTGSNACYMEEMR	213–229	661–677	9.767	0.220
2FH7 A	Receptor type Y phosphatase	9 (H)	AYIATQGPL	99–107	388–396	9.796	0.068
		11 (H)	VHCSAGVRGTG	226–236	517–527	9.780	0.148
2A38 A	Titin isoform N2-B	9 (E)	ATSTAELLV	89–97	185–193	9.753	0.236
2COT A	Kinesin-like protein KIF13B	8 (T)	VKCDECGK	19–26	47–54	9.745	0.239
1VYH C	Platelet activating factor acetyl-hydrolase IB	9 (E)	SRDKTIKMW	211–219	315–323	9.959	0.246

[$]The character within parenthesis represents the secondary structure (E for β-strand, T for turn and H for α-helix).
PDB, Protein Data Bank; RMSD, root mean square deviation, STAMP, structural alignment of multiple proteins

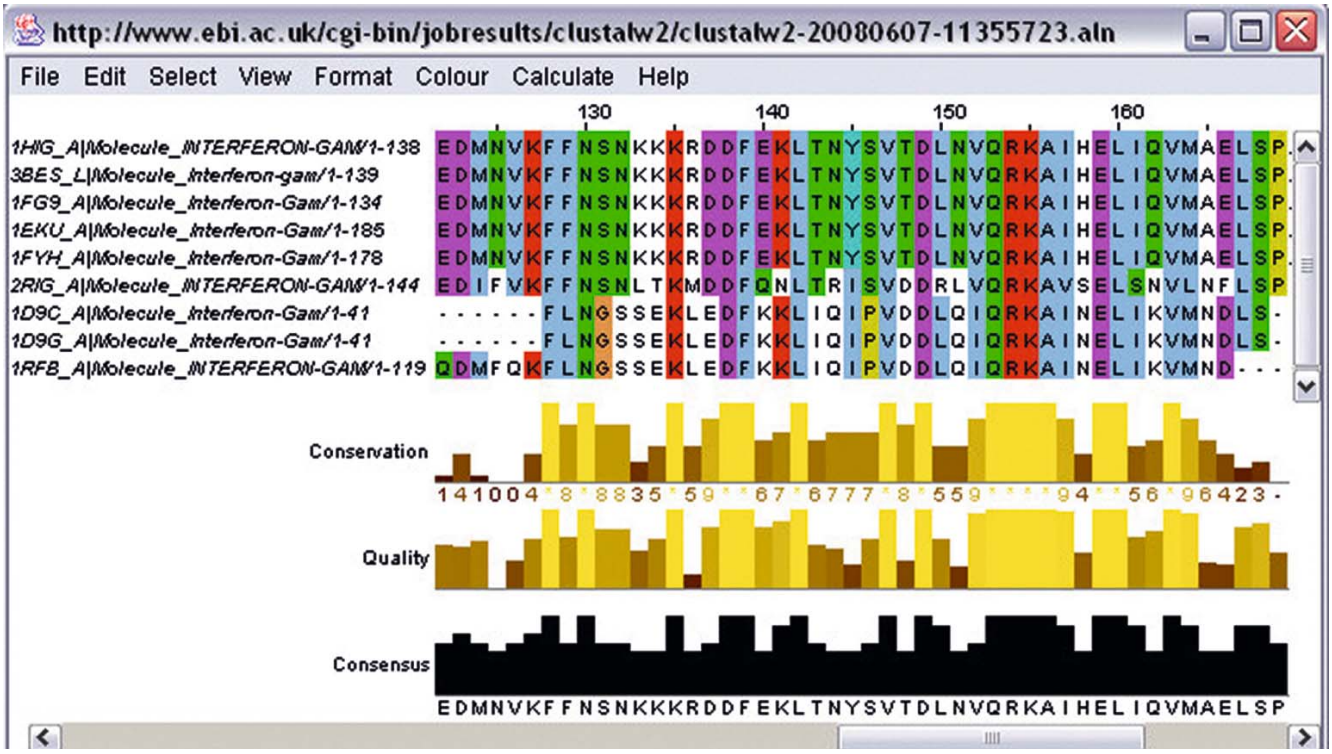

Figure 1. A screenshot of the CLUSTALW output showing the high degree of sequence conservation of interferon-γ from various species.

3.1 *Repeats in interferon-γ and hexokinases*

The repeats in interferon-γ (chains A and D of the PDB-id 1FYH), which is important in the human immune system, divide the structure into two equal and exact halves. The high degree of conservation of residues across species (figure 1) in interferon-γ as seen from the ClustalW output (Thompson *et al* 1994) leads to the conclusion that the two domains of human interferon-γ probably evolved through the process of gene duplication of the entire module rather than convergent evolution. In addition, these repeats have been conserved over a large period of evolutionary time and thus may have some useful biological function. It is worth noting that the section of the protein with known atomic coordinates contains three

Table 2. Large similar repeats from the non-redundant dataset of *Homo sapiens* proteins

PDB-id and chain	Name of protein	Repeat length[§]	Repeat	Sequence identity (%)	Location	STAMP score	RMSD (Å)
1AXN A	Annexin III	10 (H)	KAIRGIGTDE RALKGIGTDE	70	29–38 260–269	9.557	0.547
1CZA N	Hexokinase I	9 (H)	EWGAFGDDG EQGAFGDNG	88.89	260–268 708–716	9.759	0.205
		10 (H)	SGMYLGELVR SGMYLGEIVR	90	298–307 746–755	9.763	0.229
		11 (E)	ATVKMLLPTFVR AVVKMLLPSFVR	81.82	59–69 507–517	9.776	0.169
		11 (E)	GDFIALDLGGS GDFLALDLGGT	81.82	78–88 526–536	9.757	0.210
		11 (E)	GFTFSFPCQQS GFTFSFPCQQT	90.91	151–161 599–609	9.757	0.210
		11 (E)	GTGTNACYMEE GTGSNACYMEE	90.91	231–241 679–689	9.791	0.111
		11 (E/H)	TTVGVDGSLYK VTVGVDGTLYK	80.91	408–418 856–866	9.771	0.214
1E07 A	Carcinoembryonic antigen	8	LNVLYGPD LNVLYGPD LDVLYGPD	100 87.5	197–204 375–382 553–560	9.776 9.794	0.183 0.089
		8	VKTITVSA VKSITVSA	87.5	457–464 635–642	9.682	0.403
1EQF A	RNA polymerase II transcription factor	8 (H)	PMDLQTLR PMDLETIR	75	66–73 189–196	9.757	0.249
1FNH A	Fibronectin	9 (E)	TITGLQPPGT TITGLEPPGT	88.89	149–157 239–247	9.653	0.418
1FYH A	Interferon-γ	107 (H)	**Y**VKEAENLKKYFNAGHSDVADNGTLFLG ILKNWKEESDRKIMQSQIVSFYFKLFKNF KDDQSIQKSVETIKEDMNVKFFNSNKKK RDDFEKLTNYSVTDLNVQRKAIFVKEAE NLKKYFNAGHSDVADNGTLFLGILKNWK EESDRKIMQSQIVSFYFKLFKNFKDDQSIQ KSVETIKEDMNVKFFNSNKKKRDDFEKL TNYSVTDLNVQRKAI*	99.9	5–111 129–235	-	-
1H88 C	Myoglobin protooncogene	8 (H)	WTREEDEK WTKEEDQR	62.50	9–16 61–68	9.737	0.297
1KI0 A	Angiostatin	8 (E)	PWCYTTDP PWCFTTDP	57.50	63–70 144–151	9.781	0.139
1LAR A	LAR protein	12 (H)	VVHCSAGVGRTG TVHCSAGVGRTG	91.67	213–224 504–515	9.735	0.278
1M4K A	A killer cell Ig-like receptor 2DS1	8 (E)	GTYRCYGS GTYRCFGS	87.50	75–82 173–185	9.755	0.238
1M8W A	Pumilio I	8 (H)	YGCRVIQK YGCRVIQR	87.50	106–113 178–185	9.747	0.225
1M9I A	Annexin VI	10 (H)	LIEILASRTN LIEILATRTN	90.00	115–124 458–467	9.766	0.201
		10 (H)	RIMVSRSELD RIMVSRSEID	90.00	275–284 623–632	9.726	0.307
1MOX A	Epidermal growth factor receptor	8 (E)	ENLQIIRG ENLEIIRG	87.50	78–85 397–404	9.723	0.331
1N11 A	Ankyrin	8 (H)	GFTPLHVA GYTPLHVA	87.50	146–153 311–318	9.670	0.468
		9 (H)	TPLHIAARE TPLHIAAKQ	77.78	115–123 214–222	9.740	0.283

Table 2. (*Continued*)

PDB-id and chain	Name of protein	Repeat length[$]	Repeat	Sequence identity (%)	Location	STAMP score	RMSD (Å)
1NKR A	P68-CL42 KIR	8 (E)	GTYRCYGS GTYRCFGS	87.50	76–83 174–181	9.746	0.288
1O6S A	Internalin A	9 (E)	DITPLANLT DLTPLANLT	88.89	104–112 367–375	9.668	0.430
		9 (H)	NQLTDITPL NQISDITPL NQISNISPL NQISDLTPL	100 77.78 55.56 66.67	78–86 209–217 275–283 363–371	10 9.736 9.680 9.506	0.000 0.284 0.383 0.566
		10 (H)	NQISDLTPLA NQISNISPLA	70	363–373 275–284	9.713	0.334
		11 (H)	SNNQLTDITPL TNNQISDITPL	72.73	76–86 207–217	9.699	0.343
1OI1 A	SCML2 protein	11 (E)	NDFKVGMKLEA NNFKVGMKLEA	90.91	39–49 148–158	9.745	0.274
1OZ2 A	Lethal brain tumour-like protein	8 (E)	RLRLHFDG RIKIHFDG	62.50	71–78 282–289	9.712	0.308
1OZN A	Reticulon 4 receptor	9 (E)	IFLHGNRIS LFLHGNRIS	88.89	37–45 158–166	9.778	0.114
1P22 A	1P22A WD repeat protein 1A	8 (E)	DNTIKIWD DNTIRLWD	62.50	152–159 315–322	9.731	0.331
1RGO A	Butyrate response factor-2	8 (E/H)	RYKTELCR KYKTELCR	87.50	3–10 41–48	9.757	0.267
1USE A	Vasodilator-stimulated phosphoprotein	10 (H)	LQRVKQELLE LQKVKEEIIE	60.00	9–18 24–33	9.789	0.114
1YGR A	CD45 protein tyrosine phosphatase	8 (E)	LPYDYNRV IPYDYNRV	87.50	64–71 355–362	9.659	0.482
2B8L A	β-secretase 1	8 (H)	LVDTGSSN IVDSGTTN	50	50–57 246–253	9.554	0.600
2FH7 A	Receptor type protein phosphatase	8 (H)	MVQTEDQY MVQTEDEY	87.50	270–277 561–568	9.789	0.114
2H14 A	WD repeat protein	8 (E)	DDKTLKIW NDKTIKIW	62.50	90–97 306–313	8.672	1.245
		8 (E)	SGKCLKTL TGKCLKTL	87.50	101–108 143–150	9.729	0.318
2HYN A	Cardiac phospholamban	8 (H)	CLILICLL CLILICII	62.50	36–43 41–48	9.710	0.348
2ID5 A	Leucine rich repeat neuronal 6A	8 (E)	NLFNLRTL DLYNLKSL	50	78–85 126–133	9.783	0.159
2NQ3 A	Itchy homolog E3 ubiquitin protein ligase	8 (E)	EVTVDGQS EVVTNGET	–	62–69 165–172	–	–
2NZT A	Hexokinase II	8 (E)	VKMLPTFV VKMLPTYV	82.50	47–54 495–502	9.705	0.195
		11 (H)	NMEWGAFGDDG NMEWGAFGDND	90.91	244–254 692–702	9.775	0.189
		16 (H)	VAVVNDTVGTMMTCGY VAVVNDTVGTMMTCGF	93.75	190–205 638–653	9.653	0.514

$The character within parenthesis represents the secondary structure (E for β-strand and H for α-helix).

* The corresponding similar repeat has the first residue (Y) substituted by F.

PDB, Protein Data Bank; RMSD, root mean square deviation: STAMP, structural alignment of multiple proteins

repeats of lengths 87, 13 and 9 residues, respectively, which occur twice in the polypeptide chain. Interestingly, all three repeats superimpose very well and are located in the 'cup' of the Y-shaped interferon- γ, which interacts with other proteins. For example, figure 2(a) shows the superposition of the nine residues (ELIQVMAEL). These residues occur in two places (113–121 and 237–245, table 1) in the same chain of the PDB-id: 1FYH (chain A). Thus, the amino acids in the repeat are likely to be implicated in protein–protein interactions and may be exactly symmetrical to the binding site of interferon-γ.

The two identical and seven similar repeats found in Hexokinase I (PDB-id: 1CZA) and the seven identical and three similar repeats in Hexokinase II (PDB-id: 2NZT) are found in the same position (Selvarani *et al* 2004) in the two structures. Figure 2(b) shows the superposition of the eight residues (VKMLPTFV and VKMLPTYV). These similar repeats occur in two places (47–54 and 495–502, table 2) in the same chain of the PDB-id: 2NZT (chain A). It is interesting that these repeats have been highly conserved even though the domains have a sequence similarity of between 56% and 73.67%. Furthermore, these repeats surround the binding pocket of adenosine diphosphate (ADP) (figure 3) and are implicated in ADP binding. In fact, from the interactions computed using PSAP (Balamurugan *et al* 2007), the repeats either directly interact with the bound ADP or are in a position to provide scaffolding to the interacting repeats. The residues involved in adenosine triphosphate (ATP) binding – Aspartate 532 and Threonine 680 (Zeng *et al* 1996) – are a part of the repeats. It is interesting that the hexose-binding sequence LGFTFS, where Leucine is crucial for the binding of hexose (Schrich and Wilson 1987), is central to the repeats in the catalytic domains and Leucine does not form a part of the repeat in the regulatory domain of Hexokinase I. Thus, it can be concluded that the repeats in both hexokinases are conserved since they are involved in the function of the protein.

3.2 *Different amino acids have varying propensity to form repeats*

The occurrence of the allowed mutation in similar repeats is summarized in table 3. More information on the mutational preferences of residues involved in similar repeats is given in table 4. From these two tables, it can be clearly seen that of the allowed mutations, F Y is the most preferred mutation followed closely by K R. The other allowed mutations are found to occur less frequently. This pattern of occurrence suggests that certain amino acids have a greater tendency to be involved in repeats than others. Assuming that the mutation of a residue at one point is independent of the mutation at another and the rate of mutation of amino acids (or genes) remains constant over time, the propensity of an

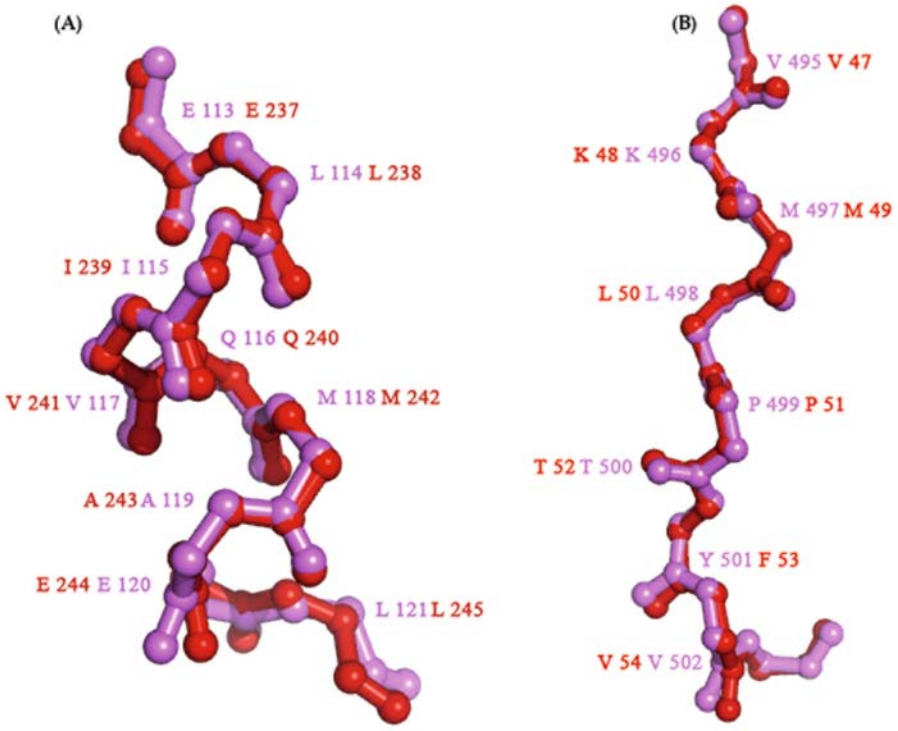

Figure 2. Superposition of the (**a**) identical and (**b**) similar internal repeats occurs in interferon-γ and Hexokinase-II structures, respectively.

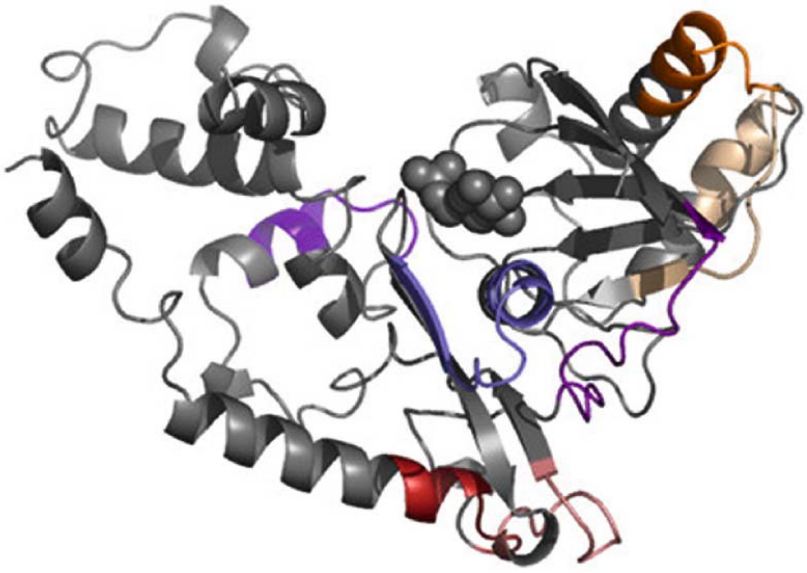

Figure 3. Repeats from the hexose-binding C terminal domain from Hexokinase-II structure are shown as coloured ribbons and the bound hexose molecule is shown as grey spheres. The corresponding repeats in the N terminal domain are not shown.

Table 3. Allowed mismatch pairs and their occurrences in similar repeats of varying length

No. of residues in the repeat	8	9	10	11	12	16	107	All
No. of proteins containing repeats	20	5	5	4	1	1	1	30*
No. of repeats	22	6	6	8	1	1	1	45
F Y	6	1	-	1	-	1	1	11
S T	7	2	3	7	-	-	-	19
V T	2	-	-	1	1	-	-	4
L I	8	5	6	2	-	-	-	21
K R	13	-	3	-	-	-	-	16
D N	4	1	1	3	-	-	-	9
Q E	8	1	1	-	-	-	-	10

*The number of proteins containing repeats of varying length is not additive because a single protein may contain multiple similar repeats.

amino acid to form repeats is given by the formula:

$$p = \sum \left(\frac{aa_{rep}}{aa_{res}} \right),$$

where p is the propensity, aa_{rep} is the number of a certain amino acid in repeats in proteins from a dataset, and aa_{res} is the number of that amino acid in all the proteins of that dataset.

The values of the propensity for each amino acid in identical and similar repeats were plotted in a graph (figure 4). The propensity of any of the amino acids (average propensity) to occur in an identical repeat was found to be 18.09% and the propensity to occur in a similar repeat was 8.89%. Interestingly, all the individual amino acids had a lower propensity to form similar repeats than identical repeats. Thus, it may be concluded that strong constraints exist to limit the mutation of amino acids in identical repeats (even through allowed mutations). Contrary to expectation, the allowed mutation pairs (called mismatch pairs) in similar repeats did not have propensity values close to each other. However, the difference in propensity between mismatch pairs was higher in identical repeats (7%) than similar repeats (2%). The propensity difference for the negatively charged/polar amino acids was the least, only around 1% (Aspartate and Asparagine had values of 10.56% and 11.87%, respectively, while the Glutamate and Glutamine pair had percentage propensities of 7.25% and 6.26%, respectively). The positively charged pair, Lysine and Arginine, with propensities of 9.88% and 6.60%, respectively, showed a greater difference than the acidic residues. Similarly,

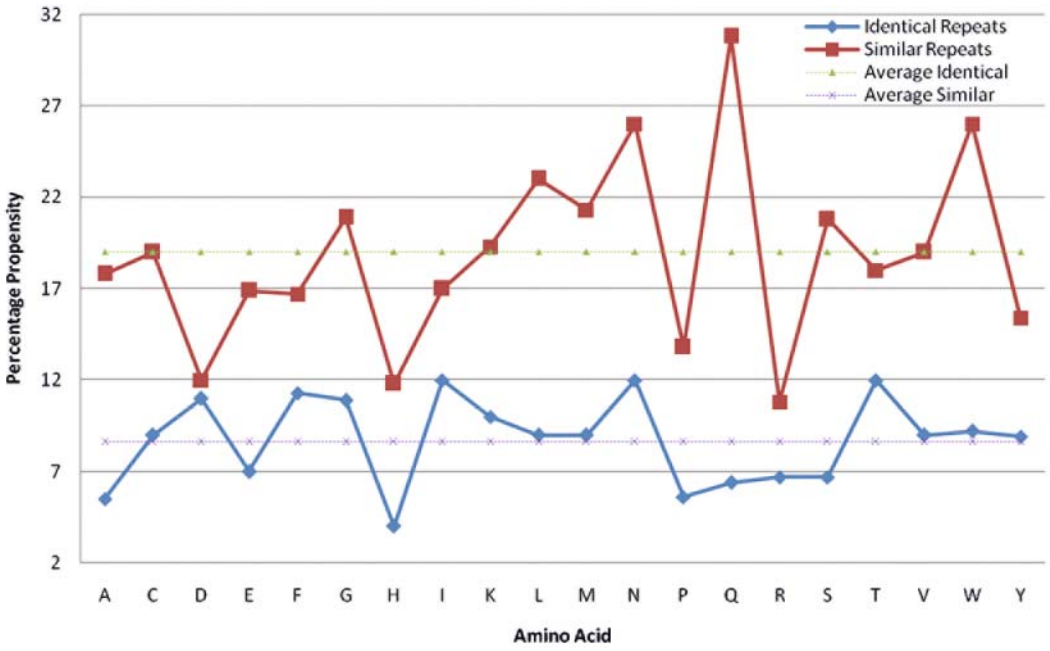

Figure 4. The propensity of different amino acids to form different types of repeats is represented as a graph. The average occurrence of all amino acids in identical or similar repeats is given as a horizontal dotted line.

the non-polar pair, Leucine, and Isoleucine had propensities of 8.84% and 11.63%, respectively. The aromatic residue pair, Phenylalanine and Tyrosine, had a median deviation in propensities of 11.00% and 9.26%, respectively.

The interconvertible trio of Serine, Threonine and Valine bear close examination. As expected from the fact that Threonine is common to both these mutations, there was a higher propensity for Threonine (11.54%) than the other two (6.56% and 9.25%) to occur in similar repeats, which had the highest difference in propensity out of all the allowed mutations. This difference can probably be explained by the fact that Threonine, acting as a link between Serine and Valine, gets accumulated in the repeats, leading to a higher propensity. Another interesting observation from the similar repeats was that Histidine, Proline and Alanine had the lowest propensities of all the amino acids. Glycine, which is very similar to Alanine, had one of the highest propensity values (10.65%). The cause for this may be the ubiquitous nature of the Glycine residue. It is also surprising that Cystine and Methionine, which are comparatively rare amino acids, had a relatively high propensity. This might imply that when these amino acids occur, they tend to be in positions that can be duplicated. Tryptophan also had a relatively higher propensity and it is intriguing that out of the seven low-propensity amino acids, four (including one allowed pair) were part of the allowed mutations. As if to balance them out, the paired amino acid had a high propensity. Thus, it would be interesting to see the percentage of occurrence of both the amino acids of the allowed substitutions (table 4). As can be seen from this table, it is obvious that values of the propensities are echoed in these mutation values.

Table 4. Allowed mismatches and their mutational preferences

Mismatch pair	No. of mutations	No. of residues of mismatches in proteins with similar repeats	Mutational preferences of allowed mismatches*
F Y	11	836	1.31%
S T V	23	3051	0.75%
L I	21	1770	1.1%
K R	16	1338	1.19%
D N	9	1275	0.71%
Q E	20	1283	0.78%

*Mutational preference = N_m*2/N_{mpr},
where N_m is the number of mismatches and N_{mpr} is the number of (mismatch pair) residues in proteins with similar repeats.

The difference between the amino acids comprising the mismatch pairs was most distinct in the identical repeats. The amine forms of the negatively charged amino acids, along with the tryptophan residue, had the highest propensity. In fact, in contrast to the similar repeats, almost half the amino acids involved in identical repeats had a propensity higher than the average. Like similar repeats, Proline and Histidine had a low propensity. This may be because Histidine is involved in the active site and is highly conserved, while Proline has a specific function in the three-dimensional structure of the protein.

4. Conclusion

We took an initial step towards understanding the constraints in the conservation of amino acid sequences by analysing large cryptic identical and similar repeats. The present study indicates that the correlation between sequence, structure and function of protein molecules can be elucidated by a careful investigation of sequence repeats. In fact, the repeats in Hexokinases and interferon-γ are probably conserved at the sequence level due to their participation in the function of the protein. Furthermore, a study of the propensity of amino acids to form repeats clearly shows that Asparagine and Glutamine are the most likely to be found in identical repeats while Isoleucine and Asparagine have the highest tendency to be found in similar repeats. In addition, it can be concluded that the most probable transition is that between the aromatic amino acids (F Y) followed by that of the positively charged pair (K R). Further work would include the expansion of this study to include a universally applicable dataset, evolutionarily related repeats (or distant repeats), small (three to five residues) and moderate-sized (five to eight residues) repeats. It would also be interesting to carry out an analysis of the positional importance of amino acids in repeats. It is hoped that this work will eventually lead to the development of a structural matrix relevant to protein structures, similar to the Point Accepted Mutations (PAM) Matrix for protein sequences.

Acknowledgements

The authors gratefully acknowledge the use of the Bioinformatics Centre (DIC), the Interactive Graphics Based Molecular Modeling facility (IGBMM) and the Supercomputer Education and Research Centre (SERC). The authors thank the Department of Biotechnology, New Delhi for funding this project. Part of this work is supported by the Department of Biotechnology-sponsored Institutewide computational biology program.

References

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

Andrade M A and Bork P 1995 Heat repeats in the Huntington's disease protein; *Nat. Genet.* **11** 115–116

Andrade M A, Perez-Iratxeta C and Ponting C P 2001 Protein repeats: structure, functions and evolution; *J. Struct. Biol.* **134** 117–131

Balamurugan B, Roshan M N A M, Hameed B S, Sumathi K, SenthilKumar R, Udayakumar A, Babu K H V, Kalaivani M, Sowmiya G, Sivasankari P, Saravanan S, Ranjani C V, Gopalakrishnan K, Selvakumar K N, Jaikumar M, Brindha T, Michael D and Sekar K 2007 PSAP: protein structure analysis package; *J. Appl. Crystalogr.* **40** 773–777

Banerjee N, Chidambarathanu N, Daliah M, Balakrishnan N and Sekar K 2008 An algorithm to find all identical internal sequence repeats; *Curr. Sci*. **95** 188–195

Djian P 1998 Evolution of simple repeats in DNA and their relation to human disease; *Cell* **94** 155–160

Hancock J M and Simon M 2005 Simple sequence repeats in proteins and their significance for network evolution; *Gene* **345** 113–118

Hancock J M, Worthey E A and Santibanez-Koref M F 2001 A role for selection in regulating the evolutionary emergence of disease-causing and other coding CAG repeats in human and mice; *Mol. Biol. Evol.* **18** 1014–1023

Heringa J 1998 Detection of internal repeats: how common are they?; *Curr. Opin. Struct. Biol.* **8** 338–345

Liou Y C, Tocilj A, Davies P L and Jia Z 2000 Mimicry of ice structure by surface hydroxyls and water of a beta-helix antifreeze protein; *Nature (London)* **406** 322–324

Marcotte E M, Pellegrini M, Yeates T O and Eisenberg D 1999 A census of protein repeats; *J. Mol. Biol.* **293** 151–160

Schrich D M and Wilson J E 1987 Rat brain hexokinase: amino acid sequence at the substrate hexose binding site is homologous to that of yeast hexokinase; *Arch. Biochem. Biophys.* **257** 1–12

Selvarani P, Shanthi V, Rajesh C K and Saravanan S 2004 BSDD: Biomolecules segment display devise – a web based interactive display tool; *Nucleic Acids Res.* **32** W645–W648

Smith T F, Gaitatzes C G, Saxena K and Neer E J 1999 The WD-repeat: a common architecture for diverse functions; *Trends Biochem. Sci.* **24** 181–185

Sumathi K, Ananthalakshmi P, Roshan M N A M and Sekar K 2006 3dss: 3-dimensional structural superposition; *Nucleic Acids Res.* **34** W128–W134

Tantz D, Trick D and Dover G A 1986 Cryptic simplicity in DNA is a major source of genetic variation; *Nature (London)* **322** 652–656

Thompson J D, Higgins D G and Gibson T J 1994 CLUSTAL W: improving the sensitivity of progressive multiple sequence alignments through sequence weighting, position specific gap penalties and weight matrix choice; *Nucleic Acids Res.* **22** 4673–4680

Zeng C, Aleshin A E, Hardie J B, Harrison R W and Fromm H J 1996 ATP-binding site of human brain hexokinase as studied by molecular modeling and site-directed mutagenesis; *Biochemistry* **35** 157–164

MS received 16 June 2008; accepted 12 December 2008

*e*Publication: 29 January 2009

Corresponding editor: VIDYANAND NANJUNDIAH