doi:10.1016/j.jmb.2003.09.023 J. Mol. Biol. (2003) 333, 1045–1059

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Structural Context of Exons in Protein Domains: Implications for Protein Modelling and Design

Bruno Contreras-Moreira†, Pall F. Jonsson† and Paul A. Bates*

Biomolecular Modelling Laboratory‡, Cancer Research UK London Research Institute Lincoln's Inn Fields Laboratories, 44 Lincoln's Inn Fields, London WC2A 3PX UK

*Corresponding author

Intron boundaries were extracted from genomic data and mapped onto single-domain human and murine protein structures taken from the Protein Data Bank. A first analysis of this set of proteins shows that intron boundaries prefer to be in non-regular secondary structure elements, while avoiding α-helices and β-strands. This fact alone suggests an evolutionary model in which introns are constrained by protein structure, particularly by tertiary structure contacts. In addition, *in silico* recombination experiments of a subset of these proteins together with their homologues, including those in different species, show that introns have a tendency to occur away from artificial crossover hot spots. Altogether, these findings support a model in which genes can preferentially harbour introns in less constrained regions of the protein fold they code for. In the light of these findings, we discuss some implications for protein modelling and design.

Keywords: protein evolution; intron–exon boundaries; comparative modelling; protein design

Introduction

Much effort has been dedicated over the last 25 years toward understanding the evolutionary meaning of introns, which were discovered independently in viral genes by the teams of Phillip Sharp and Richard Roberts.[1,2] Introns, pieces of DNA with no apparent purpose sitting inside genes, were later found to be common in eukaryotes, missing from prokaryotes but present in some archaebacteria. Two working hypotheses have been put forward in order to explain them: the introns-early and the introns-late theories.[3,4] Conservation studies across species support the existence of early introns. Non-conserved introns could be either old or recent; however, they are more likely to be acquired recently.[5,6] Evidence from protein structure analysis on a few proteins has been used by supporters of both theories to support their models,[7–9] so it is still not clear whether these theories are complementary or contradictory, though they seem to be simultaneously possible.[10] Introns can separate complete functional domains, as commonly accepted in current theories of protein evolution,[11] but they can also split the exonic components of individual functional domains, as considered here. Regardless of their origin, introns must be spliced from their mRNAs for proteins to be translated. Splicing relies on very short RNA motifs marking the boundaries; changes in these motifs affect the outcome of the splicing process directly.[12–14] Introns are recognised as genomic regions involved in insertion, deletion or duplication of new exons, or even in the formation of chimeric proteins.[15] For this reason, and this is one the main assumptions of this work, introns are potential places for insertion or deletion of fragments in proteins, and thus possible locations for significant changes in protein structure.

One of the major challenges in protein science is to understand, predict[16,17] and even design new protein functions.[18] Here, we investigate whether intron–exon boundary (IEB) information could potentially be useful in protein design, by highlighting regions within folds that are particularly resistant to natural recombination events. It is now well established that rational protein design involves searching vast sequence and conformational spaces.[19] Indeed many of the early design attempts have focused only on redesigning proteins with a fixed backbone, or allowing small

† B.C.-M. & P.F.J. contributed equally to this work.

‡ http://www.bmm.icnet.uk

Abbreviation used: IEB, intron–exon boundary.

E-mail address of the corresponding author: paul.bates@cancer.org.uk

backbone movements.[20,21] If significant modifications of functions are to be accomplished, probably much larger backbone movements will be needed. The question then arises as to how to accommodate these large changes whilst keeping the protein fold stable. A powerful approach could be to use an ensemble of homologous proteins and identify key hybridization points. Indeed, recent experimental work in this direction has been conducted.[22] Following this lead, which suggests that IEBs could lie at special locations within protein folds, we considered two completed eukaryotic genomes, mouse and man, and looked at the protein structure level to see if IEBs are special. We investigated the occurrence of introns in the context of protein secondary and tertiary structure, using a large set of human and murine protein structures, for which there is now a reasonable sample size in the Protein Data Bank (PDB).[23] By applying statistical analysis, comparative modelling and *in silico* protein recombination we obtain data to support the idea that the occurrence of introns in genes is restricted by their effect on protein structure, regardless of their evolutionary conservation. These findings could be used to improve current methods for comparative modelling and protocols for protein design.

Results and Discussion

A set of 684 single-domain human and mouse protein structures, and their amino acid sequences, extracted from the PDB (see Materials and Methods) was taken as the sample for the following statistical analysis.

Residues at intronic boundaries have a tendency to form more coil regions and fewer helices and strands than expected

Residues at IEBs were assigned a secondary structure type as calculated by the DSSP program.[24] A simple analysis was done to compare the secondary structure nature at the boundaries to the expected frequency of secondary structure states on the same dataset. The results, shown in the left-hand side of Table 1, show a significant preference for IEBs to occur in coil regions of proteins and less inside α-helices and extended β-strand elements. This could indicate that insertion of introns into sections of ordered structure, such as α-helices and β-sheets, is likely to affect the overall structure and function which, in return, affects the protein's fitness in natural selection terms. Also, even when boundaries occur within strands and helices, they tend to be close to the end of their secondary structure element, as shown in Figure 1. This is especially apparent for non-conserved IEBs in extended strands (Figure 1(A)). This supports the idea that boundaries occur in less-ordered areas.

Could the observed secondary structure biases reflect different intron types? Introns appearing in proteins as a result of late exon duplications and insertions have a phase class that is identical with that of the recipient intron.[25] An analysis of phase classes of exons and their boundaries (see the right-hand side of Table 1) does not indicate any correlation between the phasing of exons in our dataset and the secondary structure of IEBs. This, however, does not imply that phases are not conserved in particular genes, since we are comparing many different proteins from different genes. Splice variants within proteins[15] could potentially show a correlation with secondary structure at IEBs; however, insufficient data were available in our dataset for statistical analysis (see Materials and Methods).

By looking at pairs of neighbouring IEBs, instead of a single IEB, it is possible to observe biases towards certain patterns of secondary structures, which are likely to be a result of evolution, possibly through intron loss/gain events. Table 2 shows frequencies of adjacent pairs of secondary structure element at IEBs as they appear when looking at the protein sequences from the N terminus through to the C terminus. From these data, it would appear that nature favours secondary structure expansion of strands from the N terminus, rather than the C terminus. Helices appear to expand/contract with a higher frequency, with less bias towards either end.

Local structural variability at intron–exon boundaries

The relationship between structure conservation and IEBs was studied by mapping the boundaries on pairs of homologous human and mouse PDB structures with a pairwise sequence identity $\geq 40\%$. These pairs were structurally aligned and structural deviations at boundary positions compared to the overall deviation between each (see Materials and Methods). The structure conservation of boundaries in coil regions, helices and strands was not found to differ significantly from the expected values (data not shown). The location of boundaries does not therefore appear to be in significantly more structurally divergent regions between homologous proteins. Hence, the reason why these boundaries are found preferentially in coils and at the ends of α-helices and β-strands is not clear. Perhaps this is to allow variable packing of exons. To assess this, we compared the packing of exons in homologous proteins.

Packing of exons using structural alignments

We used a method based on structural alignments to assess whether exons can have alternative packing arrangements with hinge points located on IEBs. For this study, the previously described set of homologous human–mouse sequence pairs was used. Each pair was aligned by sequence and two adjacent windows, representing two exons of

Table 1. Observed and expected frequencies of secondary structure elements at intron–exon boundaries and exon phase frequencies

DSSP secondary structure, three-state structure	$f(\text{obs})_{\text{introns}}$	$f(\text{exp})_{\text{introns}}$	Difference (%)	Phase 0	Phase 1	Phase 2
C						
Not in a secondary structure element (loops)	**776 (32)**	**544 (22)**	**+43**	279	262	235
Residue in isolated β-bridge	29 (1)	31 (1)	−6	10	9	10
Hydrogen bonded turn	308 (13)	288 (12)	+7	106	111	91
Bend	260 (11)	265 (11)	−2	90	72	98
E						
Extended β-strand	**430 (18)**	**537 (22)**	**−20**	130	148	152
H						
α-Helix	**570 (23)**	**702 (29)**	**−19**	199	174	197
3_{10} Helix	73 (3)	80 (3)	−9	27	22	24
5-Helix	1 (0)	0 (0)	–	0	1	0

Coil, extended strand and helical structures are identified by the letters C, E and H, respectively. The total number of IEB residues is 2447, out of a total of 116,740 residues. The frequencies, *f*, are given as totals (%). The most statistically significant differences are highlighted in bold. The observed differences between observed and expected frequency of the secondary structure elements are highly unlikely to be random, according to a χ^2 test with six degrees of freedom ($p \ll 0.001$).

average length, were shifted along the sequence pairs, and a structural alignment performed by superimposing the two left-hand exons on each other, carrying over the structure of the right-hand exons, as described in Materials and Methods. Flexibility at each position was assessed as the angle between vectors from the N terminus to the centre of geometry of each of the right-hand exons (see the inset in Figure 2). This angle was used as an indication of the structural deviation between the pair at each point. No significant difference ($p = 0.62$ for a χ^2 test, with 12 degrees of freedom) was found in the distribution of angles at IEBs compared to background distribution as shown in

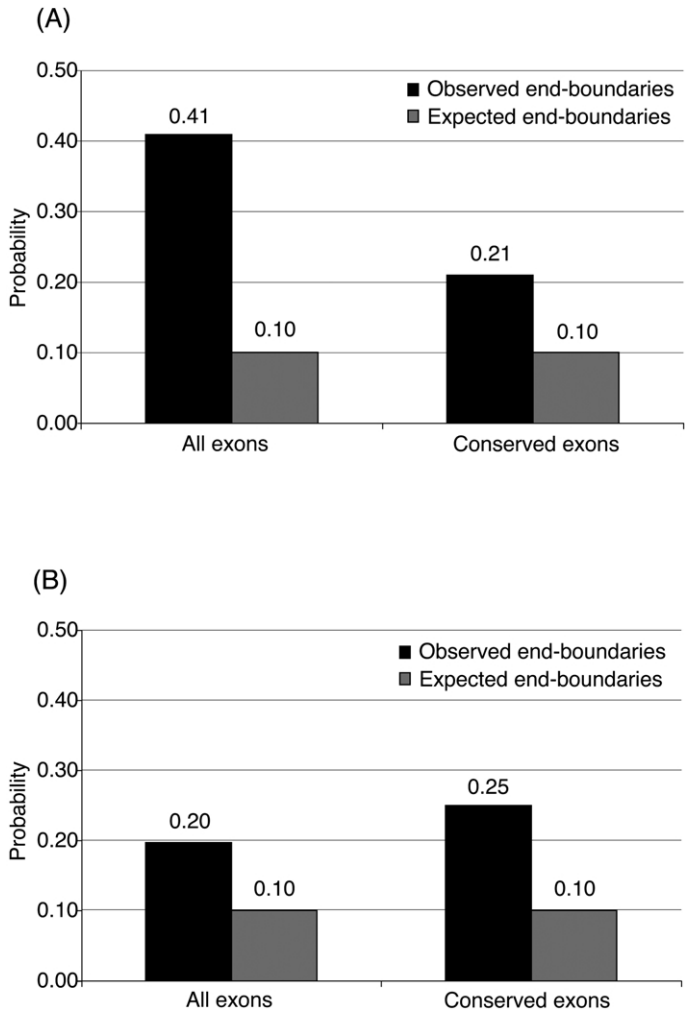

Figure 1. Frequency of intron–exon boundaries appearing at the ends of extended β-strands (A) and α-helical structures (B). Ends are defined as the first or last 5% of the secondary structure element length. Black columns show the observed frequency of boundaries in ends of secondary structure elements and grey columns show the expected frequency. Shown are the frequencies for all exons as well as a subset of conserved exons between mouse and human. The differences are significant according to χ^2 tests with one degree of freedom ($p \ll 0.001$ for all exons in extended strands and helices, $N = 450$ and 579, respectively, and $p < 0.005$ for conserved exons in extended strands and helices, $N = 62$ and 60, respectively.)

Table 2. Observed and expected frequencies of adjacent pairs of secondary structure elements on intron–exon boundaries as encountered when going from N- to C termini of sequences

IEB pairs	Observed	Expected	Difference (%)
Strand–coil	146	173	−16
Coil–strand	239	173	+38
Helix–coil	201	260	−23
Coil–helix	243	260	−7
Helix–strand	48	81	−41
Strand–helix	72	81	−11
Coil–coil	558	553	+1
Helix–helix	193	122	+58
Strand–strand	63	54	+17

Expected values were obtained by calculating the probability of occurrence of each possible pair of secondary structure states, using data in Table 1.

Figure 2. This would suggest either that evolution does not favour increased diversity of packing between homologous exons or that the method we used is not sensitive enough to pick up hinge points in boundary locations.

Analysis of tertiary structure contacts

Previous work suggests the importance of tertiary contacts in understanding the interactions between components of a fold.[22,26,27] Trying to understand our findings, we looked at the distribution of contacts around IEBs as compared to non-boundary residues along the primary sequence. Much work has been done in the past to address the conservation of introns by building multiple alignments of homologous sequences from different organisms.[5,6,28] Despite the limitation of using only human and murine proteins, we wanted to check if conserved and non-conserved introns are different in terms of contacts. The results (Figure 3(A)) show that, in our relatively large dataset, boundary residues are, in general, no different, in terms of their contact profile, compared with the rest of the protein. Low-contact regions are occupied preferably by coil residues, irrespective of the existence of a boundary there. However, as shown in Figure 3(B) and (C), coil boundary residues seem to be preferred for low-contact regions in the subset of conserved boundaries.

Location of exons in relation to functional sites

Details of functional residues of proteins in our dataset were extracted from the PDB and the spatial relationship between exons and functional sites examined. A total of 94 functional sites (as defined in PDB SITE records) were obtained from 68 PDB structures (listed in Table 3). From the total of 308 IEBs contained in this subset, 18% (55/308) are located in the vicinity (distance <7 Å) of the functional site. Similar proportions are obtained when the same calculation is repeated on sets of 308 randomly sampled residues,

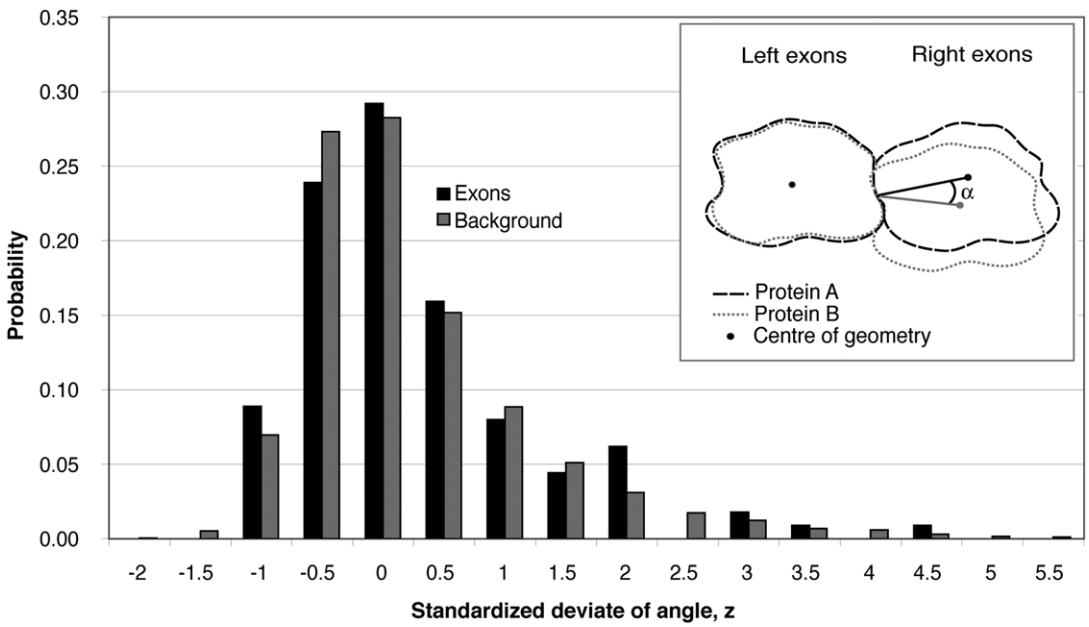

Figure 2. Distribution of standardized normal deviates of angles in intron–exon boundaries (black) and the background (grey) with a mean value of 8.11 and standard deviation of 6.76. Greater z values represent higher degree of variability between a homologous pair at a specific position. There is not a significant difference between the samples ($p = 0.62$ for χ^2_{11}). The inset shows a diagram of the calculation on a pair of proteins consisting of two exons. The centres of geometry are depicted. By superimposing the left-hand exons and carrying over the right-hand exons as rigid bodies, an angle (α) can be measured.

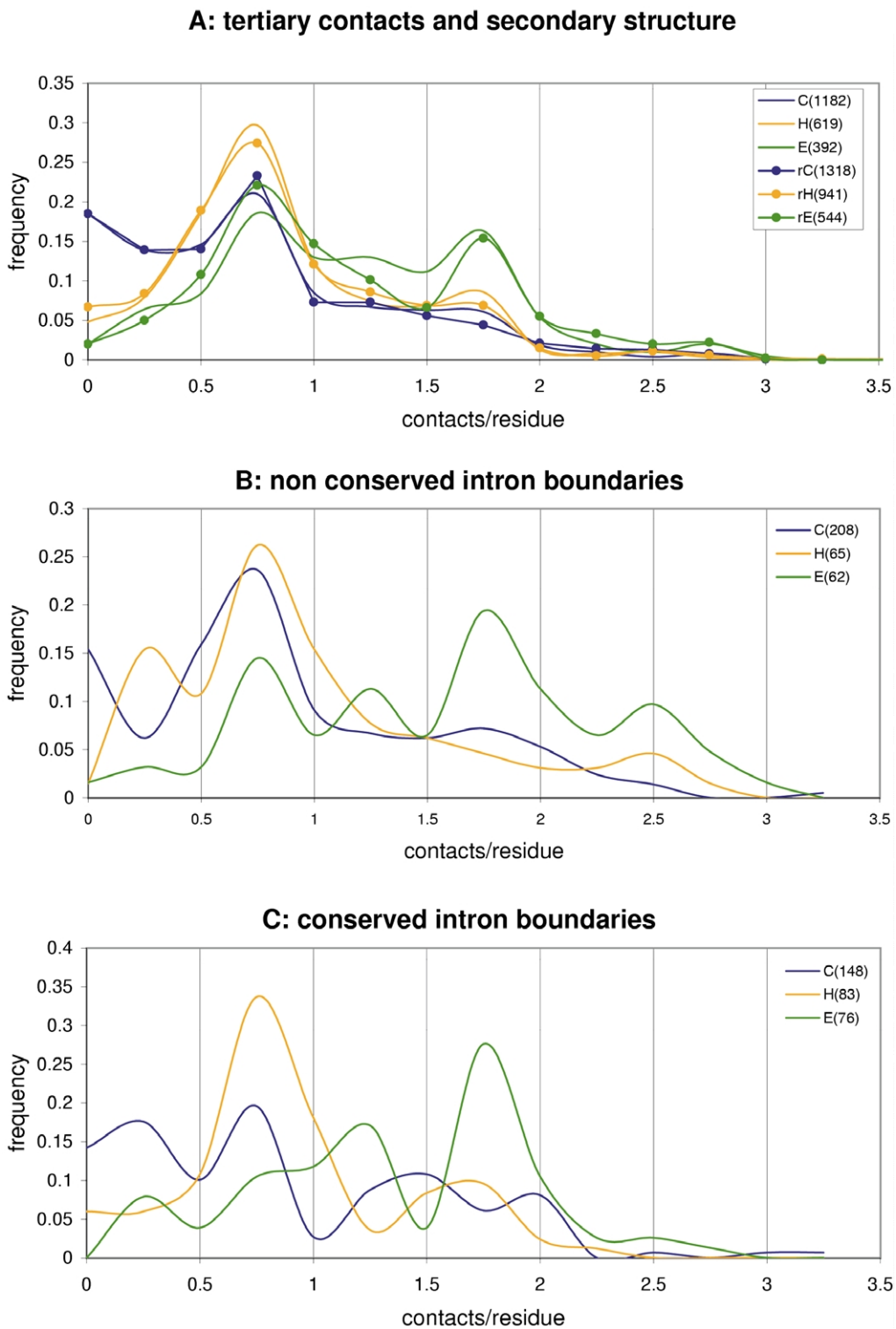

Figure 3. (A) Distribution of contacts per residue in a population of intron–exon boundaries as compared to a population of randomly chosen residues. Contacts are calculated as explained in Materials and Methods, by checking residues to the right of the selected position (intron or randomly selected) of the sequence with residues to the left. The original distribution of contacts along each sequence is smoothed by averaging with a window of size 5. Three different distributions are plotted, according to the three-state secondary structure of the selected position, where C corresponds to coil conformations, H to helices and E to extended strands (see Table 1). Random residues are labelled rC, rH and rE. The number of observations is shown in parentheses. (B) and (C) Distribution of contacts for non-conserved and conserved intron–exon boundaries for a set of non-redundant homologous pairs of human and murine proteins. These distributions were smoothed as explained above.

Table 3. Description of the 94 functional sites used in this work, as extracted from the PDB

PDB	Site	Residues	PDB annotation
1cffa	CA1	5	Calmodulin
1cffa	CA2	5	
1cffa	CA3	5	
1cffa	CA4	5	
3ayka	ZNA	3	Matrix metalloproteinase
3ayka	ZNB	3	
3ayka	CAB	3	
3ayka	CGS	12	
1gs4a	AC1	11	Human androgen receptor, ligand-binding domain (cortisol)
1gs4a	AC2	5	
1rpma	ATE	1	Protein tyrosine phosphatase *mu*
2gmfa	REA	14	Human granulocyte macrophage colony stimulating factor
1gula	GTE	11	Glutathione transferase
1gula	HTE	8	
1h4wa	CAT	3	Structure of human trypsin IV (brain trypsin)
1h4wa	BEN	8	
1h4wa	CA	4	
1h6fa	MO6	3	tbx3, t-box transcription factor, ulnar-mammary syndrome
1h6ha	AC1	8	px domain from p40phox bound to phosphatidylinositol 3-phosphate
1mema	CAT	3	Crystal structure of cathepsin k complexed with a potent vinyl sulfone inhibitor
1vhra	RCA	11	Human vh1-related dual-specificity phosphatase
1bio	NUL	3	Human complement factor D in complex with isatoic anhydride
1gxca	TPB	5	fha domain from human chk2 kinase in complex with a synthetic phosphopeptide
1h8dh	AC1	14	Human alpha-thrombin complex with a tripeptide phosphonate inhibitor
1klt	CIC	3	PMSF-treated human chymase (serine protease)
1mfma	ZN	5	Copper,zinc superoxide dismutase
1mfma	CU	4	
1trna	CAT	3	Trypsin (EC 3.4.21.4) complexed with the inhibitor diisopropyl-fluorophosphofluoridate
1h9oa	PTR	7	Phosphatidylinositol 3-kinase, p85-alpha subunit
1kpf	HNE	3	Protein kinase pkci-1 with inhibitor
1kpf	AVE	1	
5gdsh	CAT	3	Human alpha-thrombin:hirunorm V complex
1bp3a	ZNA	2	Growth hormone-prolactin receptor complex
1bsxa	A	9	Thyroid hormone receptor beta
1c25	DSU	2	cdc25a catalytic domain
1c25	POP	7	
1hazb	CAT	3	Porcine pancreatic elastase and human beta-casomorphin-7
1qf8a	ZF1	4	Casein kinase beta subunit (1–182)
1buia	ASA	3	Microplasmin-staphylokinase complex
1fit	AVE	1	Fragile histidine triad protein(chromosomal translocation)
1fj2a	ACA	3	Human acyl protein thioesterase
1hd2a	BEZ	10	Antioxidant enzyme human peroxiredoxin 51
1hdoa	AC1	24	biliverdin-ix beta reductase:NADP complex
1qh5a	ZNA	8	Human glyoxalase ii with *S*-(*N*-hydroxy-*N*-bromophenylcarbamoyl)glutathion
1hh8a	FLC	10	Phagocyte oxidase factor
1znca	CTA	5	Human carbonic anhydrase IV(lyase)
2fha	FOX	8	Human H chain ferritin
1e42a	AC1	5	Beta2-adaptin appendage domain from clathrin adaptor ap2 (Mg)
1qnta	ACC	1	Human *O*-6-alkylguanine-DNA alkyltransferase
1qr2a	ZNA	3	Human quinone reductase type 2
1uch	CAT	4	Deubiquitinating enzyme uch-l3(cysteine protease)
2hft	VII	5	Human tissue coagulation factor
2hhma	M1	7	Human inositol monophosphatase (e.c.3.1.3.25) complex with gadolinium and sulfatehydrolase
1e9ea	TMP	10	Human thymidylate kinase (f105y) complexed with dtmp
1e9ea	ADP	12	
1sra	EF1	5	Calcium-binding protein (osteonectin)
1sra	EF2	5	
1sra	MET	3	
1eaxa	SO4	4	Matriptase, membrane-type serine protease 1
1eaxa	BEN	8	
1eaza	LBS	8	Phosphoinositol (3,4)-bisphosphate binding PH domain of tapp1
1aoxa	MGA	5	I domain from integrin alpha2-beta1
1ap6a	MNA	4	Human mitochondrial manganese superoxide dismutase
1b08a	CR1	5	Lung surfactant protein D (sugar binding)
1autc	CAT	3	Human activated protein C
1rbp	R1	9	Retinol-binding protein
1rbp	R2	7	
1ggla	LBS	5	Human cellular retinol-binding protein III
1pina	ACT	3	pin1 peptidyl-prolyl *cis–trans*isomerase from *Homo sapiens*
1gkda	BUA	4	Matrix metalloprotease MMP9 active site mutant–inhibitor complex
1gloa	CAT	3	Cys25Ser mutant of human cathepsin S

(*continued*)

Table 3 Continued

PDB	Site	Residues	PDB annotation
1icfa	ACT	2	Cathepsin l (cysteine proteinase)
1ido	MG	6	I-domain from integrin CR3, Mg^{2+} bound
1cyna	BIN	13	Cyclophilin B complexed with [d-(cholinylester)Ser8]-cyclosporin
1gmya	ACT	1	Cathepsin B complexed with dipeptidyl nitrile inhibitor
1gnua	NI	2	GABA(A) receptor associated protein gabarap
1o7ka	API	2	Human p47 PX domain complex with sulphates
1o7ka	APA	3	
1psra	HO	4	Human psoriasin (s100a7),Ca^{2+} substituted for Ho^{3+} (EF-hand protein)
1rlw	CR1	12	Calcium-phospholipid binding domain from cytosolic phospholipase A2
1rlw	CR2	5	
1rlw	CR3	8	
1rlw	CA1	1	
1rlw	CA2	1	
2mfn	RGD	3	Cell attachment modules of mouse fibronectin containing the rgd and synergy regions
2mfn	SGY	5	
1npma	ACA	3	Neuropsin, a Serine protease expressed in the limbic system
1vhh	ZN1	4	Amino-terminal domain (residues 34–195) of signalling protein *sonic hedgehog*
1eaqa	CL1	3	runx1 runt domain: structural switch and bound chloride ions modulate DNA binding
1ao5a	A	3	Mouse glandular kallikrein-13 (prorenin converting enzyme)
1glqa	GA	7	Transferase (glutathione)
1glqa	HA	5	
1gmla	AC1	2	Mouse CCT gamma apical domain(chaperone)
2znc	ZN	3	Murine carbonic anhydrase IV

The Residues column indicates the number of residues within each site.

Table 4. Subset of 22 proteins used in the recombination experiments

PDB chain	(PFAM family) and annotation	No. of templates used for recombination and sequence identity range (%)	Origin of homologous proteins (templates)
1f5xa	(PF00621) Rho GEF domain	9, 100–19	Hs; Mm
1bc9	(PF01369) Sec7 guanine-nucleotide-exchange factor domain	3, 100–37	Hs; Sc
1bci	(PF00168) C2 domain of cytosolic phospholipase A2	19, 100–20	Hs; Rn; Rr
1a66a	(PF00554) Rel homology domain, eukaryotic transcription factor	11, 100–23	Hs; Mm; Ag
1ak6	(PF00241) Cofilin/tropomyosin-type actin-binding protein	9, 100–22	Hs; Mm; Ss; Ac; Sc; At
1bv8a	(PF00207) Alpha-2-macroglobulin	3, 100–62	Hs; Pd; Rn
1b4qa	(PF00462) Glutaredoxin	10, 100–20	Hs; phage T4, Ec; Ss
1ayk	(PF00413) Matrixin, metalloprotease	15, 100–59	Hs; Ss
1cmza	(PF00615) Regulator of G protein signaling domain GAIP	7, 100–31	Hs; Rn; Bt
1gcf	(PF00041) C-terminal domain of granulocyte colony-stimulating factor receptor	10, 100–16	Mm; Oc; Hs; Oa
1blj	(PF00017) BLK SH2 domain	19, 100–51	Mm; Hs; Rous sarcoma virus; Gg
1ceea	(PF00071) Ras family, CDC42	21, 100–42	Hs; Mm; St
1etc	(PF00178) Ets-domain	14, 100–36	Mm; Hs
1df3a	(PF00061) Lipocalin/cytosolic fatty-acid binding	20, 100–16	Mm; Bt; Ss; Rn
1l3na	(PF00080) Copper/zinc superoxide dismutase	12, 100–27	Hs; St; Ec; So; Bt; Xl; Pl; Ap; Sc
1gnc	(PF00489) Interleukin-6/G-CSF/MGF family	10, 100–15	Hs; Cf
1iy3a	(PF00062) C-type lysozyme/alpha-lactalbumin family	11, 100–34	Hs; Pc; Cp; Bt; Ch; Ta; Om; G.g, Cf, Eqc; Cc
1gd5a	(PF00787) PX domain	4, 100–12	Hs; Sa; Sc
1glqa	(PF00043) Glutathione S-transferase	11, 100–16	Hs; Zm; Mm; At
1f16a	(PF00452) Apoptosis regulator proteins, Bcl-2 family	12, 100–16	Hs; Rn; Ec; Mm; Kaposi's sarcoma herpesvirus
1ig6a	(PF01388) ARID/BRIGHT DNA binding domain	7, 100–20	Hs; Dm; Ec; Sc
1h4wa	(PF00089) Trypsin	14, 100–38	Rr; Ss; Bt; Hs; Ec; Rn

Ac, Acanthamoeba castellanii; Ag, Anopheles gambiae; Ap, Actinobacillus pleuropneumoniae; At, Arabdopsis thaliana; Bt, Bos taurus; Cc, Coturnix coturnix; Cf, Canis familiaris; Ch, Capra hircus; Cp, Cavia porcellus; Dm, Drosophila melanogaster; Ec, Escherichia coli; Eqc, Equus caballus; Gg, Gallus gallus; Hs, Homo sapiens; Mm, Mus musculus; Oa, Ovis aries; Oc, Oryctolagus cuniculus; Om, Oncorhynchus mykiss; Pc, Phasianus colchicus; Pd, Paracoccus denitrificans; Pl, Photobacterium leiognathi; Rn, *Rattus norvegicus*; Rr, *Rattus rattus*; Sa, *Staphylococcus aureus*; Sc, *Saccharomyces cerevisiae*; Ss, *Sus scrufa*; St, Salmonella typhimurium; So, *Spinacea oleracea;*Ta, *Tachyglossus aculeatus*, Xl, *Xenopus laevis*; Zm, *Zea mays*.

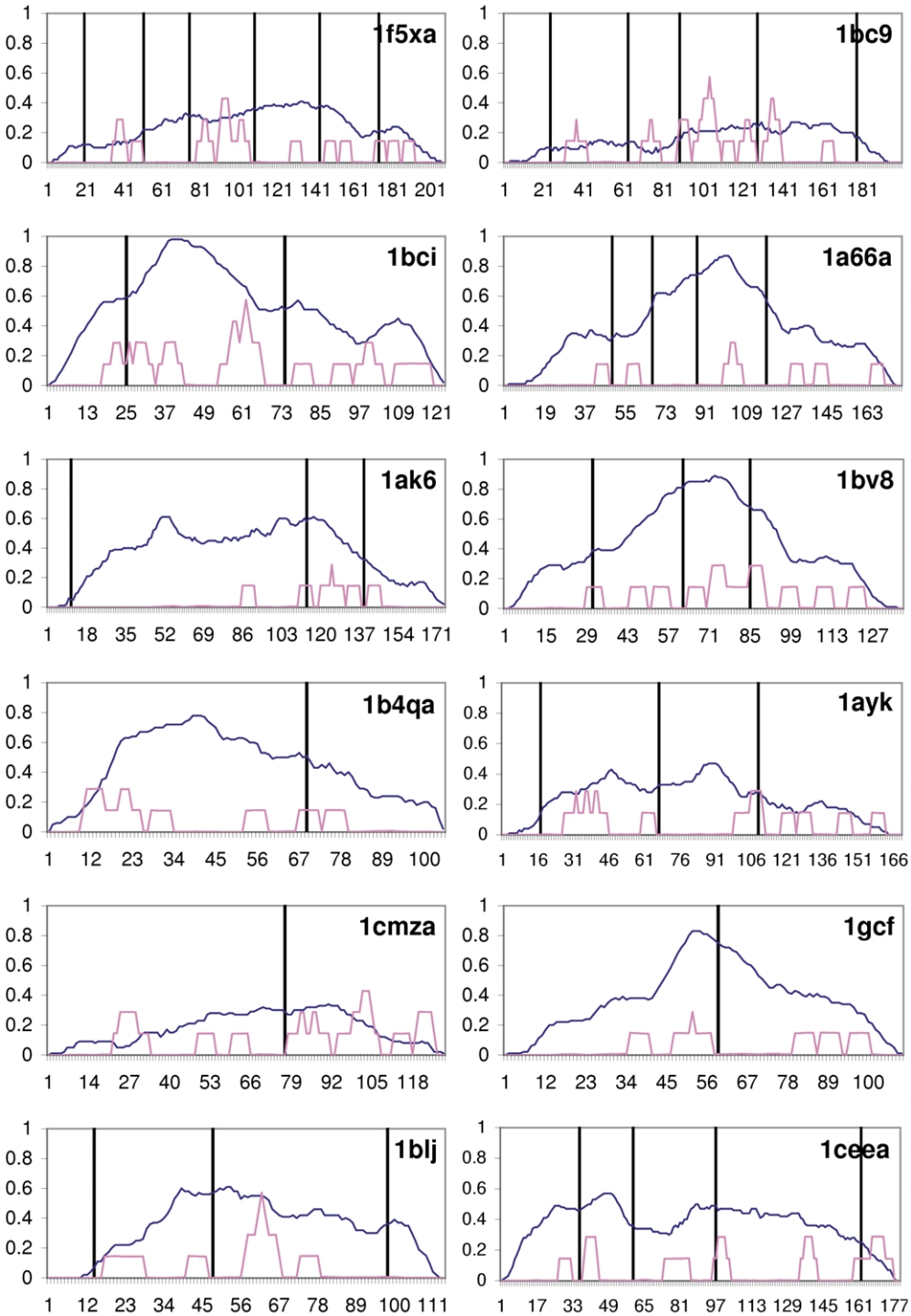

Figure 4 (*legend opposite*)

suggesting that, on average, there is no special preference for IEBs to be near important functional sites. When examining the exon composition of functional sites, we found that 34% (106/308) of intron boundaries in our set separate residues forming these sites. In total, 48 out of 94 functional sites contain residues belonging to separate exons. Again these observations follow proportions similar to those obtained when repeating the calculations with randomly chosen residues, suggesting that this is not an exclusive feature of intron boundaries. In summary, these results suggest that the pressure of selection that boundary residues support, in relation to their effect on the protein's function, is not different from that of the rest of the protein.

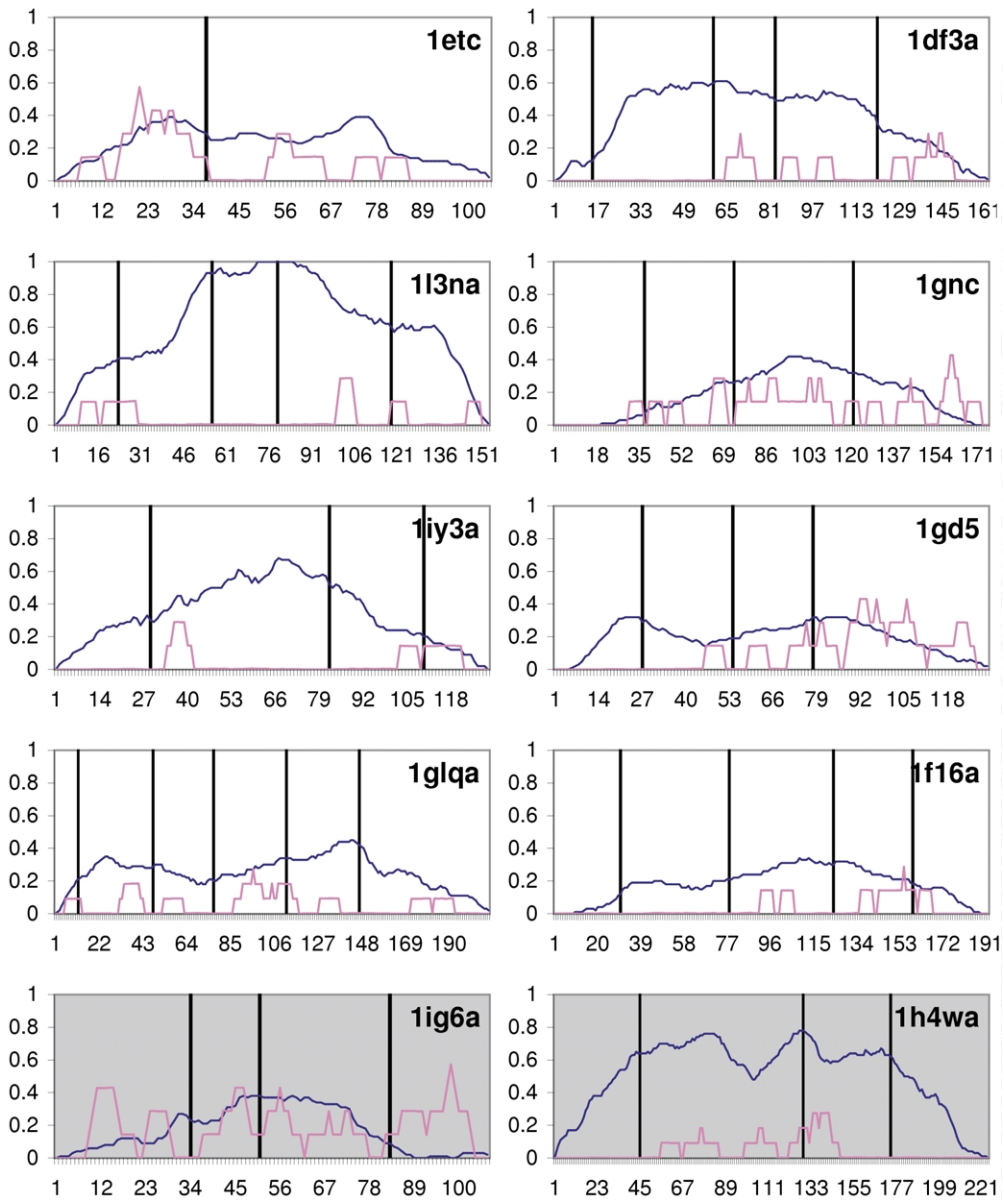

Figure 4. Frequency of crossover (pink) and tertiary contacts (blue) along the primary sequence of 22 proteins from human and mouse. Two examples explained in the text are shaded. Vertical bars indicate where natural intron–exon boundaries are found in the human or mouse sampled proteins. Crossover frequencies were smoothed by averaging inside a window of length 7 (similar plots are obtained with other values). The *Y*-axis shows the observed frequency of crossover in each of the evolving protein populations and the number of contacts divided by the length of the protein. The *X*-axis represents the amino acid sequence of each protein. Contacts are calculated as explained in Materials and Methods.

In silico recombination of protein models derived from homologous members of the same family show that crossover points tend to avoid exonic boundaries

Taken together, the results presented so far suggest that there is some evolutionary feedback between where introns reside in genes and the proteins coded by those genes, although this might have only weak connections to protein function. From an evolutionary point of view, proteins have a better chance of surviving intron loss, insertion or modification if they are in flexible or loosely packed parts of a fold, because that way the risk of disrupting the protein structure is minimized.[15,29,30] Therefore, it should be possible to find places inside particular protein folds where introns are more likely to occur. Put in a different way, introns could be marking places along a fold's primary structure, and the corresponding gene structure, where it is easier to modify proteins while maintaining the fold. However, as seen in the previous sections, contacts or flexibility alone are not enough to identify these

positions. To explore how these boundaries could be located, the following experiment was carried out.

A group of 22 human and murine proteins, extracted from the initial PDB dataset, was selected as explained in Materials and Methods. For each of them, comparative models were built using as many templates from the same or different species as possible. This included many templates for which we had no information on intron placement, and even bacterial proteins with no introns at all. This information is summarized in Table 4. The resulting 22 populations of models were recombined. From a total number of 71 boundary residues found in the dataset, 56 (79%) have less than 5% of frequency of recombination (compared to 65% expected by chance, $p = 0.01$ for χ^2_1). In other words, the observed crossover hot-spots in the 22 recombinant populations of proteins have a tendency to occur away from introns. Hence, we essentially obtain a negative image of IEB location by the use of our synthetic recombination approach. This is likely to be a consequence of the rigid crossover protocol, which is unable to emulate the natural accommodating flexibility of proteins. Because our artificial protein recombination protocol ignores where introns are and only optimizes the structural fitness of a population of proteins, these results suggest that location of introns is an important factor affecting protein fitness, in agreement with genetic evidence.[15,29,30] Indeed, Voigt and others proposed in a recent paper[22] that the correlation between introns and protein building blocks could occur as a result of natural selection, regardless of their early or late origin. However, as Figure 4 shows, contact profiles were calculated for each of the 22 populations (see Materials and Methods) and no correlation could be seen between regions with relatively few contacts and natural IEBs, as expected. This suggests that it may be too simplistic to assume that boundaries separate autonomous sections within proteins.

The fact that IEBs tend to exist away from crossover hot-spots could be applied to engineer proteins where one may want to insert fragments or to design chimeras. *In silico* recombination experiments could help in this task. In some cases, such as 1iy3a (see Figure 4), crossover regions are highly localized. Where this occurs, the information retrieved from these experiments is of little use, since large sections of the polypeptide cannot be sampled properly. In other cases, such as 1bc9, recombination hot-spots are spread along the primary structure and knowing their distribution could be a real advantage over a random guess of where intron boundaries may be located. This could be used to locate putative places for intron insertion within proteins that may have lost them, such as prokaryotic or even artificial proteins. However, it must be stated that it is not clear if the difference in the distribution of artificial and natural crossover points is a property of proteins or just a consequence of the way the recombination algorithm works. Nevertheless, the output of these simulations could be useful, especially when natural proteins show that introns can occur in any secondary structure environment and simple rules, despite the enrichment in coils observed in our data, have not been found. Two examples in which artificial recombination was applied are now explained in more detail, with the aim of illustrating the relative importance of natural and artificially selected crossover points.

Example 1: human Mrf-2 DNA-binding motif

Several structural studies on this protein[31–33] and its homologous sequences allowed us to build comparative models for all of them and perform *in silico* protein recombination, generating a profile as shown in Figure 4 (1ig6a). This protein specifically recognizes a DNA sequence through helix 5 (major groove, see H5 in Figure 5) and two loops (minor groove, L1 and major groove, L2). Note that the frequency of crossover where natural introns are contained in the gene (numbered 1, 2, 3) is low. This result could help in the task of designing a composite transcription factor by showing which regions are more spatially constrained across evolution and which are less likely to disrupt the fold if modified. In this case, the N terminal part of the L1 DNA-recognition loop is positively selected as a possible crossover point, and it is this region that is predicted to interact with DNA.[33] The C terminal part of this loop appears not to interact with DNA but it is an integral part of the fold; thus changes here could impact directly on the fold stability and hence function. On the same lines, variability could be

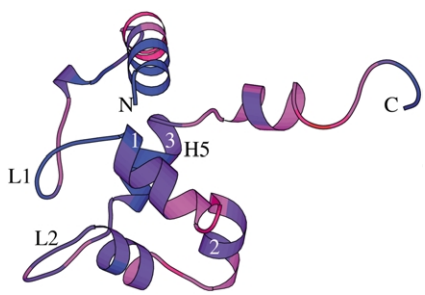

Figure 5. Protein recombination profile of human Mrf-2 DNA-binding domain mapped onto its three-dimensions model (1ig6a in Figure 4). N- and C-termini are labelled. Helix 5 (H5) and loop 2 (L2) interact with the major groove of DNA, L1 with the minor groove. Introns found in the corresponding human gene are numbered 1, 2, and 3. The frequency of recombination is mapped to the protein backbone and represented as a colour gradient. Regions close to red are positively selected as crossover points, points that anchor recombination events and improve the fitness of the fold. Blue regions were not selected in the simulation.

introduced into the major groove-recognizing helix (H5), where boundary 3 is located. However, recombining in these blue regions, e.g. near natural boundaries, could potentially cause a loss of function.

Example 2: human brain trypsin

This example was chosen because it is an enzyme containing three IEBs, marked as 1, 2 and 3 (see Figure 6(A)). Two of them are in close proximity (<7 Å) to the catalytic site, occupied in this Figure by an inhibitor, as found in the PDB.[34] A total of 14 PDB templates were used to build comparative models, with sequence identities ranging from 38% to 100%, and these were subsequently recombined (see the profile in Figure 4, 1h4wa). The frequency of crossover along the sequence is shown by the variability of the colour of the backbone in Figure 6(A). Note that most of the recorded crossover events are at the surface of the protein, away from the binding pocket, in places that, nevertheless, affect the specificity of the enzyme.[35] Unlike 1 and 3, boundary 2 is very close to a recombination hot spot and stands more than 10 Å away from the catalytic site. The four exons that comprise this protein are shown in Figure 6(B) with different colours. Clearly, the binding site is the result of the precise packing of at least three exons and thus recombining at the boundaries between these exons (1 and 3) could be directly deleterious to the protein's function.

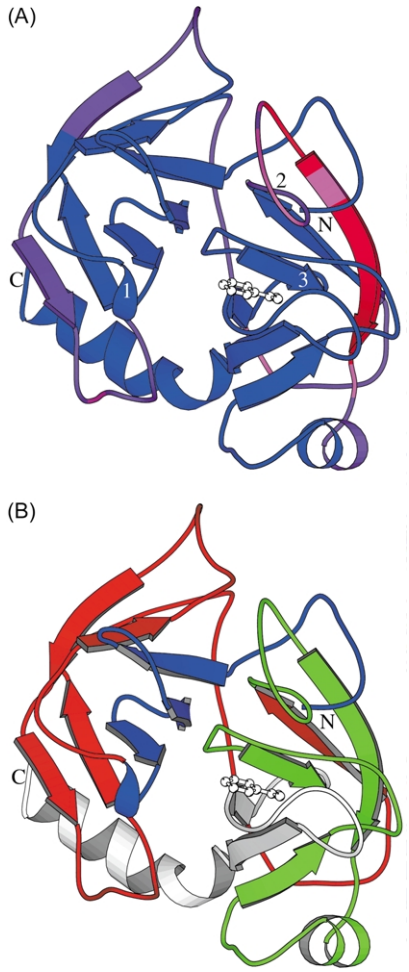

Figure 6. (A) Protein recombination profile of human trypsin mapped onto its three-dimensional model (1h4wa in Figure 4), using the same colour scheme as in Figure 5. Intron boundaries are labelled 1, 2, and 3, and the N and C termini are depicted. An inhibitor to the active site, as deposited in the PDB,[34] is shown in white. (B) Exon structure of trypsin, with four exons identified by different colours, showing that a close coordination between them is needed to form the active site.

Concluding Discussion

In higher eukaryotes, gene coding regions tend to be a small proportion of the genes; hence, there is a greater probability of natural recombination events occurring at non-coding regions, including introns. In the context of the protein fold, introns could be acting as buffer regions that accommodate exon packing upon natural recombination, or even for accommodating entirely new domains. However, in our recombination simulations we observe the opposite; crossover hot-spots steer away from intron boundaries. This might be a consequence of the superimposition-based method used for our recombination and of the complex packing between exons in some cases (as shown in Figure 6(B)). Because we treat protein fragments as rigid bodies, we cannot simulate this accommodation. Perhaps by using protein docking techniques involving some flexibility (see for example a review of current docking techniques[36]) we will be able to successfully recombine in virtually all parts of the protein, but at the cost of no longer being able to highlight natural IEBs. Thus the coarseness of our current approach may actually be an advantage.

The data presented here suggest an evolutionary feedback mechanism between natural introns and the effect they have on protein folds. Although there seems to be an enrichment of intron boundaries in coils and the ends of secondary structure elements, some natural introns occur at the midpoints of α-helices or β-strands. Therefore, in the task of designing protein recombination experiments, it is not possible to rule out regions according to their secondary structure. More complex criteria, such as protein structural fitness, tested here, may be needed. This is a stability criterion, which is not necessarily correlated to function. If function is to be modified or selected, extra restraints (or complementary functional experiments) should be required in the optimization procedure, see for example the design of a novel zinc-binding protein.[37]

The statistical analysis performed here could be useful for improving current comparative modelling protocols. In particular, after this work, it seems necessary to allow genetic algorithms to

recombine proteins regardless of the secondary structure state of the residues involved, not just in loops.[38] In addition, it could be useful to positively discriminate for intron boundary residues (when known) during recombination simulations. In these simulations, we have automatically located regions that are relatively easy to modify in structural terms, probably accounting more for the specificity of proteins rather than their function. Although predominantly within flexible regions, it seems surprisingly difficult to recombine on boundary regions, pointing to the possibility that crossing-over here may affect more dramatically protein folds and hence function. The next logical step in our analysis is to test both easy and difficult recombination examples for their effect on protein function, to be subsequentially validated experimentally.

The important question we asked ourselves at the beginning of this work was: could the knowledge of IEBs within protein families guide modelling and design? It appears from the work we have described that IEB positions may be of little use to an experimentalist wishing to design a new protein function or alter a protein's specificity by recombination; artificial recombination experiments are more likely to be successful away from IEBs. However, molecular modellers may gain insights into how algorithmic development can be facilitated. Recent CASP experiments have indicated that the field of comparative modelling, in particular, lacks a significant breakthrough in terms of more accurate algorithms.[39,40] One possible way to improve protein model building by homology is to take more account of how nature accomplishes modifications to protein structure and functions by genetic operations. Nature, it seems, can produce functional proteins with quite marked local structural disturbances upon exon remodelling.[15] We are unable to mimic many of these structural modifications *in silico*. This is probably due to our inability to sufficiently refine protein fragments upon recombination, so called cut-and-paste methods, thus this is one obvious area that must be focused upon. There are, however, other indicators of how to improve artificial recombination. For example, there appears to be a bias in neighbouring intron boundaries as to their secondary structure compositions. Such biases could be introduced into synthetic recombination.

Finally, in relation to the introns early/late debate, our findings do not allow us to exclude either theory. Some results seem to support an early origin of introns (such as secondary structure preferences), whilst others could be taken as evidence for their late origin (for example, both packing and flexibility results). Furthermore, we were not able to confidently identify older and newer introns in our dataset, since only human and mouse data were used. These results seem to agree with a model in which both theories are compatible.[10,41]

Materials and Methods

Datasets

The protein set used throughout this work was composed of human and mouse proteins obtained from the Protein Data Bank (PDB, as of 22nd January 2003).[23] To avoid large multi-domain proteins, structures with at least 100 residues but no more than 300 were selected. To avoid spliced genes, immunoglobulins and T-cell receptors were identified by sequence similarity and excluded from this dataset. Chimeric proteins were also excluded. After excluding proteins with only one exon (about 25% of the original set), this dataset contained a total of 684 PDB chains. These proteins contain, on average, 3.2 introns. Information on splicing variation of sequences in our data set was obtained from the Swissprot database (release 41.16 of 11th July 2003).[42] The number of IEBs that could be confidently assigned to splice variant parts of sequences was insufficient (42 IEBs) to allow analysis of structural distribution of IEBs in alternate exons and is not presented here. For the study of human–mouse homologous proteins, human and mouse sequence pairs of sequence identity $\geq 40\%$ were extracted from the above dataset, resulting in 118 pairs. Many homologous sequences are contained in this set but no effort was made to remove redundancy, since it was observed that almost identical proteins may have a different number of introns, in different positions along the sequence.

A subset of 22 proteins (shown in Table 4), selected to cover different folds and functions was used to perform *in silico* recombination experiments with comparative models built from close and remote homologous structures in the PDB. These 22 proteins were selected to avoid multi-domain proteins, and have diverse comparative modelling templates that could be confidently aligned.

Assignment of introns to protein sequences

Intron–exon boundaries (IEBs) were assigned by mapping protein sequences to the human (NCBI Human Contig Assembly 31, November, 2002 freeze) and murine (MGSCv3 release 3, February, 2002 freeze[43]) genome assemblies, using the BLAT server.[44] When using protein amino acid sequences in this work, introns are defined as the residues corresponding to the left-hand side of the boundary at DNA level. IEBs in homologous proteins are said to be conserved if they occupy exactly the same place in the structural alignment of those proteins. Phases of exons at IEBs were obtained by dividing the genomic position of the last DNA base of each exon by three and calculating the modulus.

Secondary structure, comparative modelling

Protein secondary structure was assigned using the program DSSP.[24] Comparative protein models were built using the server 3D-JIGSAW[45] in the interactive mode, using alignments with bit-scores of at least 1.8 and as many different templates as possible. Some templates were extracted from the corresponding PFAM families (see Table 4) using the web server DomainFishing.[46] Protein structure Figures were prepared using Rasmol[47] and Molscript.[48]

Calculation of contacts

To calculate the tertiary contacts around a given residue *r*, every C^{β} from residues to the left of *r* was checked against every C^{β} to the right in the protein sequence, calculated in a similar fashion to Voigt and co-workers.[22] A contact was then defined as a pair of C^{β} separated less than 7.0 Å in Cartesian space and more than four residues in sequence, as described.[49]

Recombination of proteins

The protein recombination protocol used is a modification of a published one[38] that adds new side-chains in every mutation event using the program SCWRL[50] and performs up to five rounds of steepest descent minimization on every newly created sibling. These simulations where executed on a 2.8 GHz PC, taking several minutes in the best case and up to 20 hours in the worst. This non-deterministic algorithm can be represented as follows:

initial population of models

↓

(1) **grow population** : r recombination + $(1 - r)$mutation

↓

(2) **select best proportion according to fitness**

↓

(3) **converged**? stop : otherwise back to (1)

Fitness is calculated with a simple potential energy function based on two terms: statistical atomic potentials extracted from the PDB[51] and solvent accessibility parameters†.[52] Crossover events (occurring with frequency *r*) along the sequence of successful models are recorded in real time (in the PDB format *B*-factor column) to be analyzed later.

Local flexibility at intron–exon boundaries

A subset (118) of homologous human–mouse pairs with pairwise sequence identity ≥40% was extracted from our original dataset. IEBs were mapped onto PDB structures and each of the human–mouse sequence pairs superimposed using MSUPER, an in-house implementation of a well-established progressive multiple structural alignment algorithm.[53,54] A window of seven residues was moved along the superposition and the fitness of the alignment recorded by summing MSUPER alignment scores, ranging from 0 for a good fit to 9 for a bad fit (C^{β}–C^{β} distances), for each of the seven positions. The DSSP program[24] was used to calculate secondary structure elements for aligned sequences and residues classified as participating in a strand, helix or coil region. The window scores for each of the three secondary structure elements were then normalised and the scores for IEBs were compared to the overall expected scores.

Packing of exons using structural alignments

The average exon length in the dataset of 118 human and mouse sequence pairs of sequence identity ≥40% was calculated. This average value (41 residues) was increased by 5% to compensate for alignment gaps between the pairs, bringing the exon length to 43. Sequence pairs were aligned using CLUSTALW[55] and pairs containing more than 20% alignment gaps were excluded. Two adjacent windows of the average exon length, representing two theoretical exons, were moved along the aligned sequence pair and a structural alignment performed using MSUPER, superimposing the two left-hand exons on each other and carrying over the structure of the right-hand exons as rigid bodies (see Figure 2 inset). A vector from the N terminus of the right-hand exon to the centre of geometry of the same exon was calculated for both sequences and the angle between the vectors determined. This was repeated for the whole length of the sequence alignment. Sequence alignments too short to yield at least 30 angles were excluded, taking the total number of pairs to 112. The data were normalised and angles for positions, where conserved IEBs occurred, were compared to the expected values.

† http://wolf.bms.umist.ac.uk/naccess/

Acknowledgements

We thank Cancer Research UK for supporting our work, and the Biomolecular Modelling Laboratory for their input and discussions.

References

1. Berget, S. M., Moore, C. & Sharp, P. A. (1977). Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. *Proc. Natl Acad. Sci. USA*, **74**, 3171–3175.
2. Chow, L. T., Gelinas, R. E., Broker, T. R. & Roberts, R. J. (1977). An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA. *Cell*, **12**, 1–8.
3. Gilbert, W. (1987). The exon theory of genes. *Cold Spring Harbor Symp. Quant. Biol.* **52**, 901–905.
4. Palmer, J. D. & Logsdon, J. M., Jr (1991). The recent origins of introns. *Curr. Opin. Genet. Dev.* **1**, 470–477.
5. Fedorov, A., Cao, X., Saxonov, S., de Souza, S. J., Roy, S. W. & Gilbert, W. (2001). Intron distribution difference for 276 ancient and 131 modern genes suggests the existence of ancient introns. *Proc. Natl Acad. Sci. USA*, **98**, 13177–13182.
6. Fedorov, A., Merican, A. F. & Gilbert, W. (2002). Large-scale comparison of intron positions among animal, plant, and fungal genes. *Proc. Natl Acad. Sci. USA*, **99**, 16128–16133.
7. Stoltzfus, A., Spencer, D. F., Zuker, M., Logsdon, J. M., Jr & Doolittle, W. F. (1994). Testing the exon theory of genes: the evidence from protein structure. *Science*, **265**, 202–207.
8. de Souza, S. J., Long, M., Schoenbach, L., Roy, S. W. & Gilbert, W. (1996). Intron positions correlate with module boundaries in ancient proteins. *Proc. Natl Acad. Sci. USA*, **93**, 14632–14636.
9. de Souza, S. J., Long, M., Schoenbach, L., Roy, S. W. & Gilbert, W. (1997). The correlation between introns and the three-dimensional structure of proteins. *Gene*, **205**, 141–144.
10. de Souza, S. J., Long, M., Klein, R. J., Roy, S., Lin, S. & Gilbert, W. (1998). Toward a resolution of the

introns early/late debate: only phase zero introns are correlated with the structure of ancient proteins. *Proc. Natl Acad. Sci. USA*, **95**, 5094–5099.
11. Chothia, C., Gough, J., Vogel, C. & Teichmann, S. A. (2003). Evolution of the protein repertoire. *Science*, **300**, 1701–1703.
12. Padgett, R. A., Grabowski, P. J., Konarska, M. M., Seiler, S. & Sharp, P. A. (1986). Splicing of messenger RNA precursors. *Annu. Rev. Biochem.* **55**, 1119–1150.
13. Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. & Watson, J. D. (1994). *Molecular Biology of the Cell*, 3rd edit., Garland, New York.
14. Clark, F. & Thanaraj, T. A. (2002). Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. *Hum. Mol. Genet.* **11**, 451–464.
15. Patthy, L. (1999). *Protein Evolution*, Blackwell Science, Oxford.
16. Devos, D. & Valencia, A. (2000). Practical limits of function prediction. *Proteins: Struct. Funct. Genet.* **41**, 98–107.
17. Thornton, J. M., Todd, A. E., Milburn, D., Borkakoti, N. & Orengo, C. A. (2000). From structure to function: approaches and limitations. *Nature Struct. Biol.* **7**, 991–994.
18. Doi, N. & Yanagawa, H. (1999). Design of generic biosensors based on green fluorescent proteins with allosteric sites by directed evolution. *FEBS Letters*, **453**, 305–307.
19. Looger, L. L. & Hellinga, H. W. (2001). Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: implications for protein design and structural genomics. *J. Mol. Biol.* **307**, 429–445.
20. Reina, J., Lacroix, E., Hobson, S. D., Fernandez-Ballester, G., Rybin, V., Schwab, M. S. *et al.* (2002). Computer-aided design of a PDZ domain to recognize new target sequences. *Nature Struct. Biol.* **9**, 621–627.
21. Looger, L. L., Dwyer, M. A., Smith, J. J. & Hellinga, H. W. (2003). Computational design of receptor and sensor proteins with novel functions. *Nature*, **423**, 185–190.
22. Voigt, C. A., Martinez, C., Wang, Z. G., Mayo, S. L. & Arnold, F. H. (2002). Protein building blocks preserved by recombination. *Nature Struct. Biol.* **9**, 553–558.
23. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al.* (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235–242.
24. Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
25. Patthy, L. (1987). Intron-dependent evolution: preferred types of exons and introns. *FEBS Letters*, **214**, 1–7.
26. Berezovsky, I. N., Grosberg, A. Y. & Trifonov, E. N. (2000). Closed loops of nearly standard size: common basic element of protein structure. *FEBS Letters*, **466**, 283–286.
27. Berezovsky, I. N. & Trifonov, E. N. (2001). Van der Waals locks: loop-n-lock structure of globular proteins. *J. Mol. Biol.* **307**, 1419–1426.
28. Betts, M. J., Guigo, R., Agarwal, P. & Russell, R. B. (2001). Exon structure conservation despite low sequence similarity: a relic of dramatic events in evolution? *EMBO J.* **20**, 5354–5360.
29. Wessler, S. R. (1989). The splicing of maize transposable elements from pre-mRNA—a minireview. *Gene*, **82**, 127–133.
30. Purugganan, M. & Wessler, S. (1992). The splicing of transposable elements and its role in intron evolution. *Genetica*, **86**, 295–303.
31. Yuan, Y. C., Whitson, R. H., Liu, Q., Itakura, K. & Chen, Y. (1998). A novel DNA-binding motif shares structural homology to DNA replication and repair nucleases and polymerases. *Nature Struct. Biol.* **5**, 959–964.
32. Whitson, R. H., Huang, T. & Itakura, K. (1999). The novel Mrf-2 DNA-binding domain recognizes a five-base core sequence through major and minor-groove contacts. *Biochem. Biophys. Res. Commun.* **258**, 326–331.
33. Zhu, L., Hu, J., Lin, D., Whitson, R., Itakura, K. & Chen, Y. (2001). Dynamics of the Mrf-2 DNA-binding domain free and in complex with DNA. *Biochemistry*, **40**, 9142–9150.
34. Katona, G., Berglund, G. I., Hajdu, J., Graf, L. & Szilagyi, L. (2002). Crystal structure reveals basis for the inhibitor resistance of human brain trypsin. *J. Mol. Biol.* **315**, 1209–1218.
35. Perona, J. J. & Craik, C. S. (1997). Evolutionary divergence of substrate specificity within the chymotrypsin-like serine protease fold. *J. Biol. Chem.* **272**, 29987–29990.
36. Janin, J., Henrick, K., Moult, J., Eyck, L. T., Sternberg, M. J. E., Vajda, S. *et al.* (2003). CAPRI: a critical assessment of predicted interactions. *Proteins: Struct. Funct. Genet.* **52**, 2–9.
37. Petersen, K. & Taylor, W. R. (2003). Modelling zinc-binding proteins with GADGET: genetic algorithm and distance geometry for exploring topology. *J. Mol. Biol.* **325**, 1039–1059.
38. Contreras-Moreira, B., Fitzjohn, P. W. & Bates, P. A. (2003). *In silico* protein recombination: enhancing template and sequence alignment selection for comparative protein modelling. *J. Mol. Biol.* **328**, 593–608.
39. Tramontano, A., Leplae, R. & Morea, V. (2001). Analysis and assessment of comparative modeling predictions in CASP4. *Proteins: Struct. Funct. Genet.* Suppl, **22–38.**
40. Tramontano, A. (2003). Of men and machines. *Nature Struct. Biol.* **10**, 87–90.
41. Fedorova, L. & Fedorov, A. (2003). Introns in gene evolution. *Genetica*, **118**, 123–131.
42. Bairoch, A. & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucl. Acids Res.* **28**, 45–48.
43. Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P. *et al.* (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
44. Kent, W. J. (2002). BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664.
45. Bates, P. A., Kelley, L. A., MacCallum, R. M. & Sternberg, M. J. (2001). Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. *Proteins: Struct. Funct. Genet.*, 39–46.
46. Contreras-Moreira, B. & Bates, P. A. (2002). Domain fishing: a first step in protein comparative modelling. *Bioinformatics*, **18**, 1141–1142.
47. Sayle, R. A. & Milner-White, E. J. (1995). RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.* **20**, 37–376.
48. Kraulis, P. J. (1991). MOLSCRIPT: a program to

produce both detailed and schematic plots of protein structures. *J. Appl. Crystallog.* **24**, 946–950.
49. Hu, J., Shen, X., Shao, Y., Bystroff, C. & Zaki, M. J. (2002). Mining protein contact maps. In *BIOKDD02: Workshop on Data Mining in Bioinformatics* (Zaki, M. J., Wang, J. T. L., Toivonen, H. T. T., eds), pp. 3–10, ACM, Edmonton, Canada.
50. Dunbrack, R. L., Jr & Karplus, M. (1993). Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J. Mol. Biol.* **230**, 543–574.
51. Robson, B. & Osguthorpe, D. J. (1979). Refined models for computer simulation of protein folding. Applications to the study of conserved secondary structure and flexible hinge points during the folding of pancreatic trypsin inhibitor. *J. Mol. Biol.* **132**, 19–51.
52. Eisenberg, D. & McLachlan, A. D. (1986). Solvation energy in protein folding and binding. *Nature*, **319**, 199–203.
53. Russell, R. B. & Barton, G. J. (1992). Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins: Struct. Funct. Genet.* **14**, 309–323.
54. Gerstein, M. & Levitt, M. (1996). Using iterative dynamic programming to obtain accurate pairwise and multiple alignments of protein structures. *Proc. Intl Conf. Intell. Syst. Mol. Biol.* **4**, 59–67.
55. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673–4680.

Edited by J. Thornton

(Received 20 June 2003; received in revised form 27 August 2003; accepted 10 September 2003)