

# Splice Variants: A Homology Modeling Approach

Nicholas Furnham,<sup>1</sup> Stuart Ruffle,<sup>1</sup> and Christopher Southan<sup>2\*</sup>

<sup>1</sup>*School of Biological and Chemical Sciences, University of Exeter, United Kingdom*

<sup>2</sup>*Oxford Glycosciences (UK) Ltd, United Kingdom*

**ABSTRACT** Splice variants play an important role within the cell in both increasing the proteome diversity and in cellular function. Splice variants are also associated with disease states and may play a role in their etiology. Information about splice variants has, until now, mostly been derived from the primary transcript or through cellular studies. In this study information from the transcript and other studies is combined with tertiary structure information derived from homology models. Through this method we have determined that it is possible to effectively model splice variants. Forty models of splice variants for fourteen proteins were produced. Analysis of the models shows that deletions produce superior model validation values. Additions to sequences where there is little homology become increasingly difficult to model with increasing sequence length. Many of the splicing events are associated with post-translational modification either in the N-terminal region by changing the signal peptide or by affecting the number or availability of glycosylation sites. Often the alternative exon combinations are associated with loss or gain of whole structural units, as opposed to just changing small loop regions. Losing part of the secondary structure may destabilize neighboring parts of the same secondary structure. Detailed analysis is given of four biomedically relevant proteins (Beta-site Amyloid Precursor Protein Cleaving enzyme (BACE), Interleukin-4, Frataxin and Hereditary hemochromatosis protein) and their associated splice variant models. The visualization of these possible structures provides new insights about their functionality and the possible etiology of associated diseases. *Proteins* 2004;54:596–608.

© 2003 Wiley-Liss, Inc.

**Key words:** alternative splicing; comparative modeling; BACE; Frataxin; Interleukin-4; hereditary hemochromatosis protein

## INTRODUCTION

Increasingly it is being realized that alternative splicing is a key mechanism for expanding proteome diversity, providing an explanation for the difference in the number of genes being discovered compared to the number of genes predicted, given the number of proteins.<sup>1</sup> The increased level of molecular diversity generated by alternative splicing can be seen in a wide range of cell types. This has been revealed particularly in neurological tissue where alterna-

tive splicing has been shown to play a crucial role in cell function such as in neurological signaling,<sup>2</sup> axon guidance,<sup>3</sup> and hair cell tuning.<sup>4</sup> In addition to this there are a number of examples where alternative splice forms exist of proteins that have associated disease states. Four examples are used in this study. First, beta-site amyloid precursor protein cleaving enzyme (BACE) which cleaves the amyloid precursor leading to the generation of the amyloid protein involved in plaque formation in Alzheimer's disease.<sup>5</sup> Second, interleukin-4 which is associated with atopic asthma.<sup>6</sup> Third, frataxin, a mitochondrial protein whose partial loss in function is associated with Friedreich's ataxia, a progressive neurodegenerative disease.<sup>7</sup> Fourth, hereditary hemochromatosis protein (HFE), associated with *porphyria cutanea tarda*, *variegates porphyria*, and hereditary hemochromatosis, the later caused by altered iron stores leading in midlife to clinical complications.<sup>8</sup>

Much of the information about the alternative exon combinations of alternative splice forms have been derived from primary sequence analysis, by the comparison of EST and mRNA sequences compared against genomic sequence. Information derived from genomic analysis has been combined with cellular studies to begin to develop an understanding of the functionality of alternative splice forms and their effects on the etiology and pathology of the associated disease states.

Here a novel approach is taken to investigating the possible functional significance of alternative splice forms by determining their putative tertiary structure by homology modeling. Although the protein databank (PDB)<sup>9</sup> includes all public protein structures, we were unable to find any examples of experimental crystal structures for alternative splice forms of the same protein in the current release of PDB. Given this and the information resolved on alternative splice forms and exon structure from the primary transcript it should be possible to derive homology models of alternative splice forms if one of the splice variants has had its structure resolved. A model of the possible splice variants can be used to link the genomic information with the proteomic information, providing a

Nicholas Furnham's present address is Department of Biochemistry, University of Cambridge, Cambridge, United Kingdom.

\*Correspondence to: Christopher Southan, Oxford Glycosciences (UK) Ltd, The Forum, 86 Milton Park, Abingdon, Oxon, OX14 4RY, UK. E-mail: chris.southan@ogs.co.uk

Received 6 February 2003; Accepted 15 June 2003

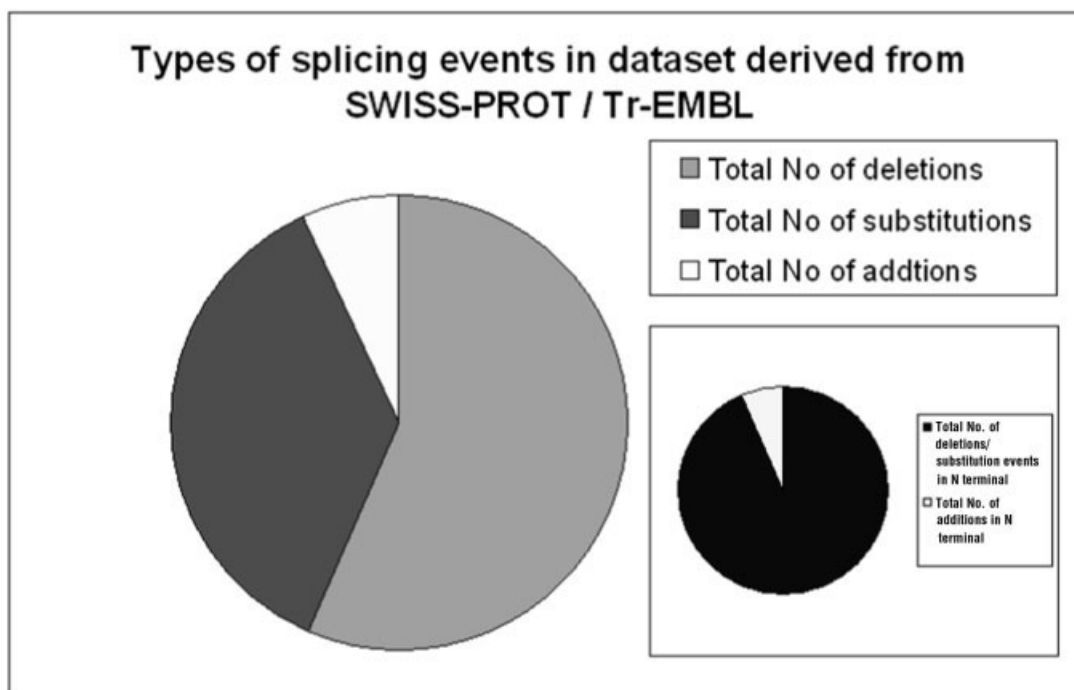


Fig. 1. Venn diagram of the proportions of splicing events for the dataset derived from SWISS-PROT where there was VARSPLIC annotation.

greater understanding of the etiology of associated disease states as well as a greater understanding of the functional significance of the splice variants.

## MATERIALS AND METHODS

Information on alternative splice forms was gathered from SWISS-PROT/Tr-EMBL database<sup>10</sup> through SRS6,<sup>11</sup> using information derived from SWISS-PROT VARSPLIC<sup>12</sup> analysis. The information was collated, using a program developed by us, into a local relational database developed in InterBase.<sup>13</sup> Further information of splicing events was derived from the local database using the SQL-92 query language. Exon information was derived from LocusLink (LocusLink at the NCBI <http://www.ncbi.nlm.nih.gov/>). This also provided a check for further alternative splice forms derived by LocusLink through mRNA and EST alignments. Structures for modeling were derived by determining the intersection of the human splicing dataset and the PDB database using SRS6. Further candidate templates were derived using the SWISS-MODEL blast tool.<sup>14</sup> Templates to be used were assessed by sequence homology. Modeling was performed with the model-default option of the program MODELLER, Version 6.0.<sup>15</sup> Sequence alignment for the homology modeling was carried out using ClustalW multiple global alignment program in BioEdit<sup>16</sup> and refined manually. Validation of the models and templates was made by PROCHECK<sup>17</sup> (with a default resolution value of 2.00Å for the splice models), PROMOTIF<sup>18</sup> and Verify3D.<sup>19</sup> Further refinement of the models was carried out in SWISS PDB viewer.<sup>20</sup>

## RESULTS AND DISCUSSION

A SRS search of the SWISS-PROT/Tr-EMBL (taken as of June 2002) database for proteins that had associated splice variant information returned 3028 entries. This equated to 6141 individual splicing events, either a sequence addition/deletion/substitution, averaging two splicing events per protein. The overall proportions of splicing events are shown in Figure 1. The majority of splicing events were deletions (56%). A significant proportion (15%) of splicing occurred in the N terminal region. This is important for the modeling process as resolved crystal structures are of the mature processed protein that are in many cases missing the N terminal region. Thus there is no homology in this region to use as a template to model against. This region is often associated with post translation modification processes. This significantly reduces the possible modeling dataset. Similar queries of splicing events were made (as a subset of the dataset) for different species. Interestingly, for *Homo sapiens* (with 2884 associated splicing events) and *Mus musculus* (with 1172 associated splicing events) the relative proportions of splicing events were similar not just to each other but to the dataset as a whole. This compares to *Drosophila melanogaster* (with 532 associated splicing events), which shows a greater amount of splicing occurring in the N terminal region. The *Homo sapiens* data subset was further analyzed to determine the size of the splicing events in the N-terminal region. Only a small proportion of the splicing events were additions (4%) with most (77%) of the events being deletions or substitutions of less than 142 residues (see Figure 2). This is as expected as much of the N-

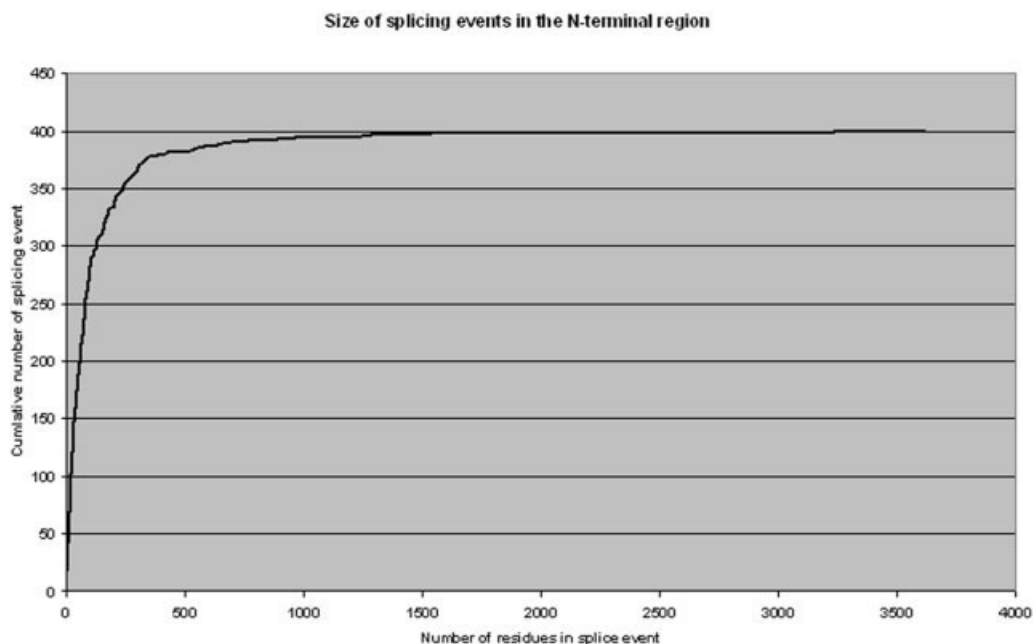


Fig. 2. Plot of the size of splicing event occurring in the N-terminal region. The N-terminal region is defined by the splicing event starting at position one of exon one.

terminal region is involved in post-translation modification such as signal peptides and proprotein domains as well as alpha helical transmembrane anchoring regions.

In determining the proteins that could be effectively modeled, the intersect of SWISS-PROT/Tr-EMBL *Homo sapiens* data subset with the PDB database was taken returning 128 entries. The PDB files were searched for sequence homology to the splice variant sequences derived from SWISS-PROT. This produced a putative modeling dataset of 14 proteins with 26 alternative splice forms. Except for RAS-related C3 botulinum toxin substrate where the splice variant is an addition, most (62%) of the splicing events in the modeling dataset are deletions with the rest being substitutions (see Table I). A summary of the modeling dataset with the PDB files used and the validation of the modeled alternative splice forms is given in Appendix A. All the models produced, except for beta-glucuronidase—which has a very poor resolved structure as its only available template—have a good or very good assessment of overall stereochemistry as assessed by the Ramachandran plot values. Residues in the most favored regions are in excess of 85% (maximum value 95.4%), with most of the other residues found in the additional allowed regions.

Below are examples of some of the isoform models that have been resolved. They have been chosen for their biomedical significance and quality of models. A summary of the template and model validation is shown in Table II.

#### Beta-site Amyloid Precursor Protein Cleaving Enzyme (BACE)

A key event in the pathogenesis of Alzheimer's disease is the cleavage of amyloid protein ( $A\beta$ ), which forms the

major component of the senile plaques in the brain characteristic of Alzheimer's disease.  $A\beta$  is formed by the proteolytic cleavage of the amyloid precursor protein by  $\beta$  and  $\gamma$ -secretases cleaving at the N- and C-terminal ends respectively. Four alternatively spliced forms of the  $\beta$ -site cleaving enzyme (BACE) have been determined.<sup>21</sup> There are nine exons that make up the long isoform encoding 501 amino acid residues. The shortest isoform, Isoform D (also referred to as BACE-I-432), is missing the last 44 residues of exon 3 and the first 25 residues of exon 4. This corresponds to a loss of two beta-strand regions and an alpha helix in a region that supports one side of the active site cleft and a possible overhang structure, consisting of two anti-parallel beta-strands orientated over the active site. The model produced of this isoform has a relatively good set of Ramachandran plot values (87.8% in most favored regions), as well as showing little deviation from the mean for the main and side chains. The first 55 residues could not be effectively modeled due to their absence from the template structure. This region corresponds to a 21-amino acid signal peptide followed by a 24-amino acid proprotein domain. Similarly the last 55 residues could not be effectively modeled, again due their absence in the template. This C-terminal region corresponds to a transmembrane domain and a 24-amino acid C-terminal cytoplasmic tail. Two glycosylation sites at position 152 and 172 (by BACE-501 sequence numbering) have been lost, leaving only two glycosylation sites at positions 223 and 354. The models produced for the other two alternative splice forms show less alteration to the active site cleft, though all three still maintain the possible overhang structure of the two anti-parallel beta-strands orientated over the active site (See Figure 3). Isoform C





(also referred to as BACE-I-457) has lost the last 44 residues from exon 3 which corresponds to two beta-strand regions and two glycosylation sites, while Isoform B (also referred to as BACE-I-476) has lost the first 25 residues of exon 4 corresponding to a short (eight residue) alpha-helix. Both have good Ramachandran values (91.8% and 90.0% respectively of residues in most favored regions). The Ramachandran values for the models are in fact better than those of the template structure. This could be an effect of the modeling procedure spatially locking the sequence with perfect homology. As the whole of the sequence is of perfect homology, the energy minimization step improves the stereochemical properties of the model when compared to the template when analyzed by PROCHECK leading to the Ramachandran plot data. This is supported by the Verify3D output (See Figure 4). Isoform B results shows the areas around the deletion site are a close match to the template structure. With the only part of the structure falling below 0.2 (which is the score indicating that residue types in a window of 21 is being found in an environment for which it has a low propensity) is in the C terminal region. This low propensity score is also seen in the template structure. In fact the area prior to this drop off has values which are better than the template. This is also seen in isoforms C and D, though the N terminal section does show a drop in propensity scores, particularly in isoform C around the splice site. Isoform D though has very good propensity values around the splice site that is better than the template.

It is interesting to note that if the full spliced out sequence (i.e. of isoform D) is compared, by a running a Blast search, against the PDB database there is a small structurally conserved sequence of 10 residues found in a number of homologs. These include Cathepsin D (1LYW chain B), Chymosin (3CMS fragment 1) and Pepsin hydrolase (1PSA chain B). The region corresponds to a beta sheet and turn. The functional significance of this is difficult to determine without experimental data. The models produced support the report of the effect of BACE isoforms C and B on A $\beta$  production by Tanahashi and co-workers<sup>20</sup> suggesting that, though both are N-glycosylated, these forms were not produced in a mature form when compared to the mature BACE-501 form produced by the Golgi. This suggests that they underwent different post-translational pathways. It was also noted that these isoforms, through measuring A $\beta$  secretion compared to BACE-501, also had reduced proteolytic activity. The model by Sauder and co-workers<sup>22</sup> of BACE-501 complexed with its substrate shows several hydrophobic contacts between the position 1 Met (after the cleavage point) of the  $\beta$ -cleavage site and BACE residues Leu19, Tyr132 and Ile179 and between the position 3 Val residue and the BACE Phe170.<sup>21</sup> The deterioration of the active site in the models, including the loss of key complexing residues Phe 170 and Ile 179, supports the view that the isoforms have effects directly on the enzyme activity.

### Interleukin-4

Interleukin-4 (IL-4) has two alternative splice forms identified: Interleukin-4 (IL-4) and Interleukin-4 $\delta$ 2 (IL-4 $\delta$ 2). IL-4 is made up from 4 exons, with exon 2 missing in IL-4 $\delta$ 2. Experimentation by Aamas and co-workers<sup>23</sup> has shown that IL-4 $\delta$ 2 is differentially expressed in the thymus and airways and inhibits the function of complete IL-4. It is also proposed that the balance between IL-4 and IL-4 $\delta$ 2 may be important in the regulation of IL-4 effects. Interestingly, a molecular model of IL-4 $\delta$ 2 (by homology modeling) has been proposed by Zav'yalov and co-workers,<sup>24</sup> created using SYBYL software with just a single template (PDB code 1RCB). It showed that the loop between the first and second helices and the first beta strand are lost without any significant changes to the hydrophobic core and the native fold including the second beta strand. This compares to the homology model built in this study where three templates were used including 1RCB, which on template validation is shown to be the poorest of the templates available. The model produced has a good quality set of Ramachandran validation values (88.8% of residues in most favored regions). It too has lost the loop region and first beta strand between the first and second helices but the presence of the second beta strand is not confirmed. Verify3D output (data not shown) demonstrates that the model is valid, though there is a drop in propensity values in the N terminal region, which is probably a result of the loss of the opposing beta-strand in the beta-sheet region. This demonstrates an important general consideration in modeling alternative splice forms. If an exon, or part of an exon, is missing then the remaining section may not have the same secondary structure as the longer form. Losing part of the secondary structure may destabilize neighboring parts of the same secondary structure. This may also be seen in terms of helix stability where a change in the number of residues within an alpha helix from odd to even may reduce stability. A similar destabilizing action can be seen over longer-range interactions with beta sheets, as with IL4, by removing a whole or part of an anti-parallel strand of the sheet. It has also been suggested by Zav'yalov and co-workers<sup>23</sup> that there is a slight rotation of the first alpha helix that changes the binding activity of IL-4 $\delta$ 2 by changing the distance of Glu33 and Lys36 (using full sequence numbering) residue side chains away from the critical orientation for the binding in the IL-4 receptor. A similar rotation is seen in this study. This, in the absence of a validation dataset within the PDB (determined by keyword searches in all annotation fields of the PDB and if a match was made comparing sequence homology), gives some demonstration that the models produced here are comparable to other published modeling attempts.

Interestingly, when the mRNA data is compared to non-human mRNA data the absence of exon 2, i.e., the presence of IL-4 $\delta$ 2, is seen more in the mRNA transcript in vertebrates than in other species (USCS Genome Browser at <http://genome.ucsc.edu/index.html>). IL-4 is a prime example of an exon deletion corresponding to a discrete structural unit the loss of which, has a subtle effect on the

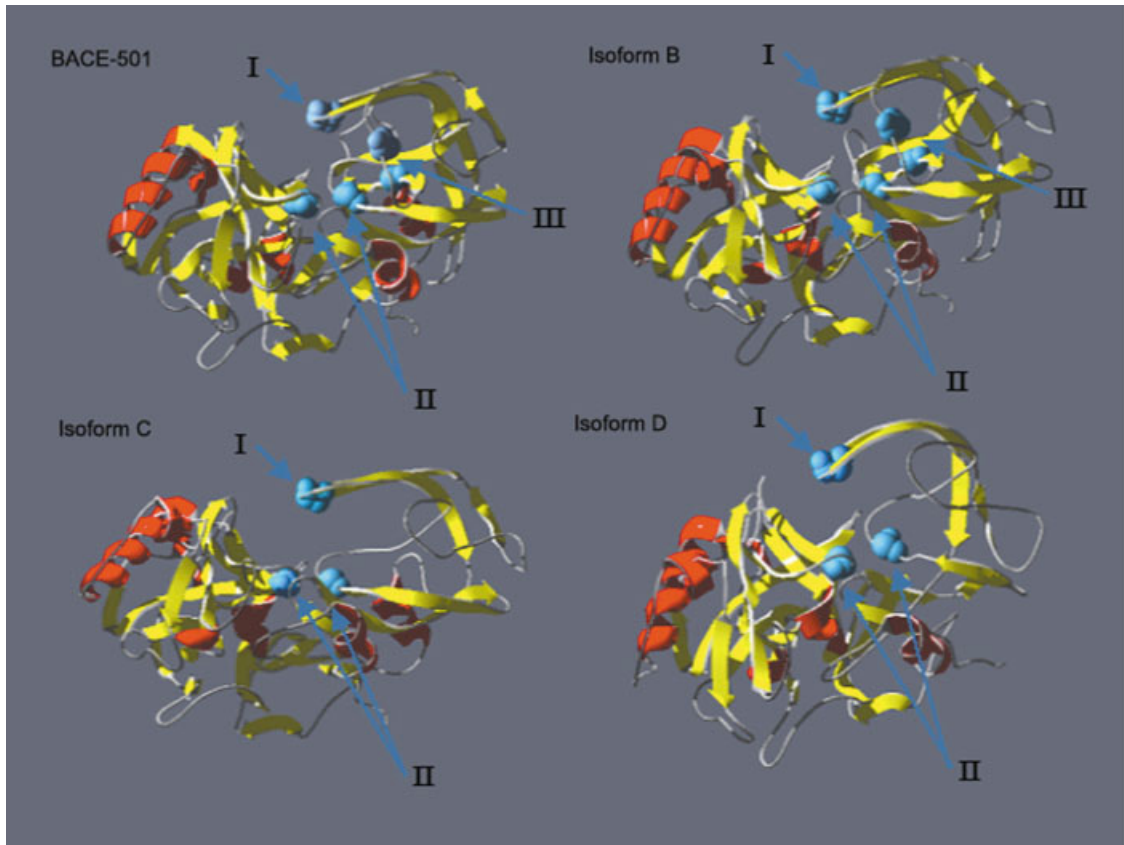


Figure 3.

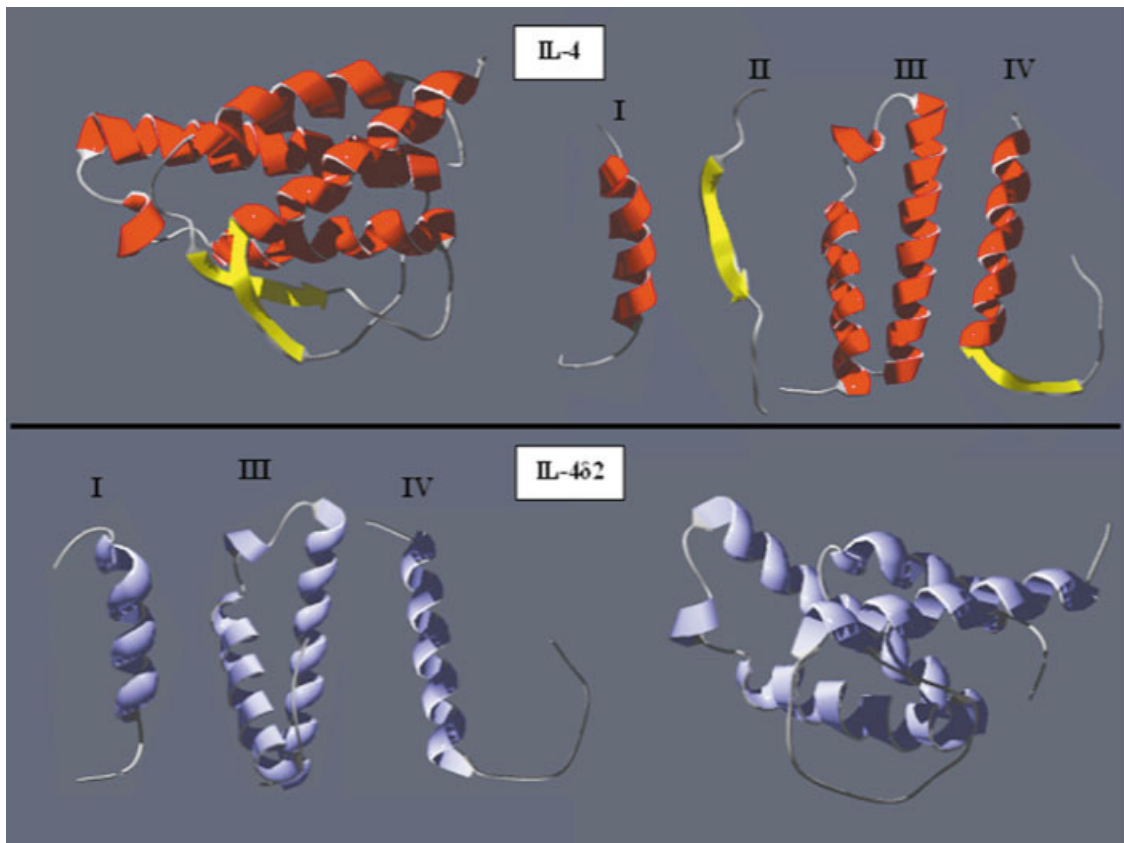


Figure 5.

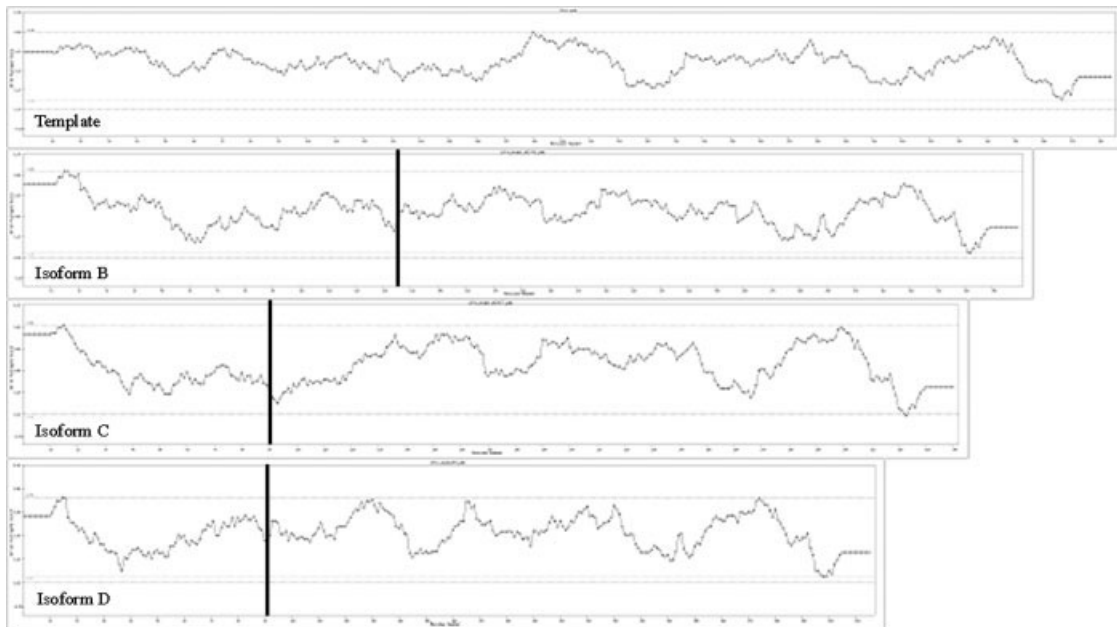


Fig. 4. Graphs of the Verify3D results for the template structure (BACE-501) and the spliced isoform models (Isoforms B, C, and D) showing the averaged environmental propensity scores for each residue. The solid vertical line on for each of the isoforms shows where the splice site is.

rest of the structure and has a profound effect on functionality (see Figure 5). This is in contrast to the more gross changes observed in BACE.

### Frataxin

Frataxin is a small mitochondrial protein. The low or lack of expression of which leads to a disorder, Friedreich's ataxia (FA), characterized by progressive neurological disability and heart abnormalities.<sup>25</sup> The disease state is an inherited recessive disorder caused by GAA triplet repeat extensions in the first intron of the gene which, by forming an unusual DNA structure prevents successful transcription. Isoform 2 where the C-terminal, exon 5a, is replaced with a short sequence of 11 amino acids. This new sequence is termed exon 5b and is located 40kb from the end of exon 5a in the telomeric direction.<sup>26</sup> Our model of this splice variant shows a loss of a beta strand, an extended U-shaped loop region followed by a short alpha

helix and a further extended loop region. The first 83 residues could not be effectively modeled due to their absence in the resolved template structures. Part of this initial N-terminal sequence is the mitochondrial transit peptide (residues 1–41). The eleven-residue exon 5b sequence was modeled by MODELLER as a floating chain. The model (as produced by MODELLER) has Verify3D output (data not shown) considerably better in the 30-residue region prior to the splice site, with the propensity values quickly dropping in the splice addition region as it was modeled as a floating chain which, by its nature, has a flexible structure. Further analysis in SWISS PDB viewer showed the possibility that if this sequence was transformed around the Pro69 (using the model numbering) the sequence lies parallel to the final beta strand of the four strand beta sheet. Though there is low sequence homology to the original beta strand, the replaced sequence seems likely to also be a beta strand. With this new conformation the Ramachandran values increased slightly. There were neither clashes between the chains or indication from PROCHECK or PROMOTIF of compromised bond angles or lengths. Using the hydrogen-bonding tool in SWISS PDB viewer (See Figure 6), there is indication of some hydrogen bonding between the new beta stand and the rest of the sheet, thereby maintaining the five-strand beta sheet.

Recently a second isoform has been identified which has an eight-base-pair insertion between exon 4 and 5a, leading to a frameshift in the mRNA reading frame thereby introducing a new stop codon in exon 5a.<sup>27</sup> This produces a 196-amino-acid protein termed isoform A1. An attempt was made to model this new splice variant, but due to the length (46 residues) of the new sequence with no sequence or structural homology available, the resulting model was

Fig. 3. Display of the experimentally resolved structure of beta-site amyloid precursor protein cleaving enzyme (BACE-501) and the models of its three splice variants (Isoform B, C, and D). The structures are shown with ribbon representation to show secondary structure elements (Helix denoted in red, strand in yellow). Also shown are key residues (in blue): I indicates the tyrosine residues at position 132 which makes a hydrophobic contact with position 1 of the  $\beta$ -cleavage site; II indicates the two aspartic acid active site residues at position 93 and 289; III indicates the two residues, phenylalanine and isoleucine at positions 170 and 179 respectively, which make hydrophobic contact with positions 1 and 3 of the  $\beta$ -cleavage site.

Fig. 5. Display of the experimentally resolved structure of interleukin-4 and the model of its splice variant interleukin-4 $\delta$ 2. The structures are shown with ribbon representation to show secondary structure elements (Helix denoted in red, strand in yellow, model structures in purple). For each the structure is split by exon (numbered I–IV). In the model of IL-4 $\delta$ 2 a slight twist to the second helix in exon III can be seen, as well as the possible loss of the second beta-sheet in exon IV.

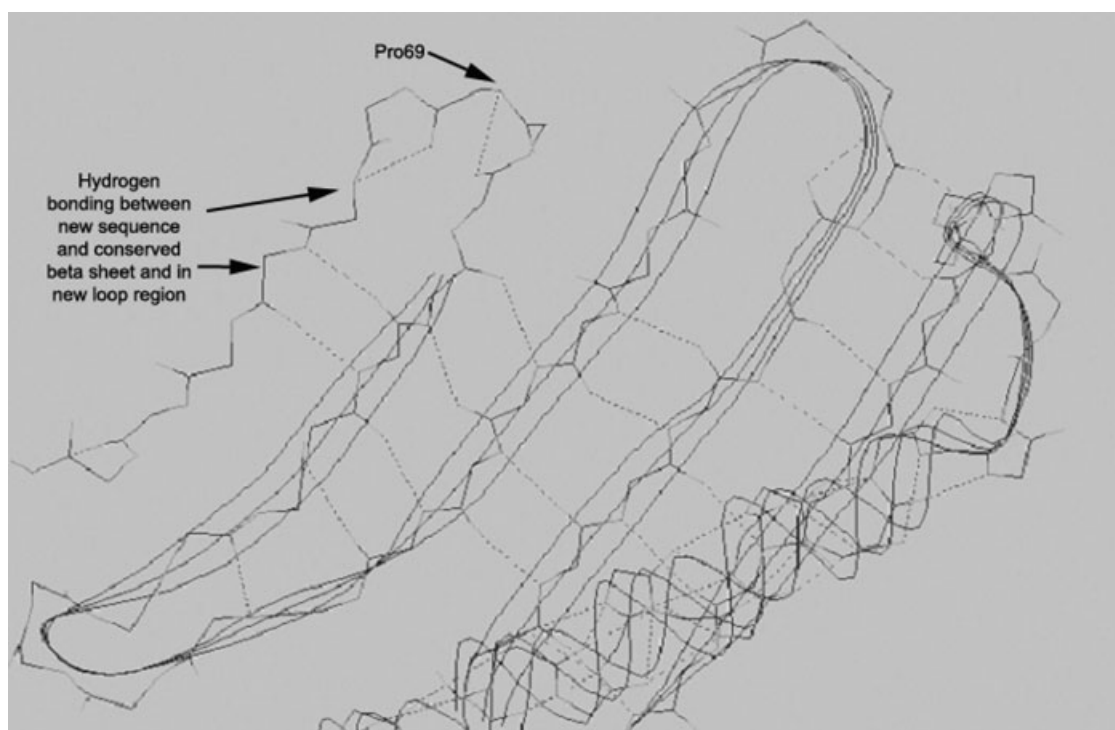


Fig. 6. Close-up of the beta-sheet region of the frataxin short isoform model. The structure is shown with non-space filled ribbon representation to show definite secondary structure elements. Through this is shown the main chain backbone (solid line) with its hydrogen bonding (broken lines).

inconclusive. The presence of low or zero homology to large sequence additions is one of the limiting factors affecting our modeling procedure. It has been shown by Pianese and co-workers<sup>26</sup> that isoform A1 is expressed in low levels in both FA patients and control individuals, while it is uncertain of the level of expression, if any, of isoform 2. Functional significance has not yet been determined for either of these isoforms.

### Hereditary Hemochromatosis Protein

Hereditary hemochromatosis protein (HFE), resembling the major histocompatibility complex class 1 molecules, is encoded by six exons. According to SWISS-PROT there are four alternative splice forms. Isoform two is missing the majority of exon two, retaining just the first 12 residues of the exon. This corresponds to a four-strand beta sheet, consisting of two long and two short anti-parallel beta strands, and a 15-residue alpha helix. Isoform 3 is missing the first 13 residues of exon four, which corresponds structurally to a beta strand from a four-stranded anti-parallel beta sheet. The fourth isoform is missing both of these regions of exon 2 and 3. As is indicated by the stereochemical analysis there is a wide range of results for the validation of the models, though most seem to be an improvement on the template structures. This is again probably a factor of the modelling process as the Verify3D data (data not shown) shows a wide range regions where there is an increase (compared to the template) of residue types with higher propensity scores and regions which fall below the 0.2 score optimal minimum. These regions do

tend to correspond to regions in the template that also fall below the 0.2 score optimal minimum. A further seven possible alternative splice forms that have been identified by mRNA alignment in NCBI LocusLink. It is important to note that the NCBI isoform 2 is not the same as the SWISS-PROT isoform 2; the SWISS-PROT isoforms 2, 3, and 4 are in fact equivalent to NCBI isoforms 7, 6, and 8 respectively. There are three known glycosylation sites, one in each of exons two, three and four. The models of the isoforms produced here do not include the first 26 residues, which make up the signal peptide of the N-terminal region, or the last 50 residues in the C-terminal region - corresponding to a transmembrane region (equivalent to exon five) and a cytoplasmic tail (equivalent to exon six), due to the lack of resolved structure in the templates. NCBI isoforms 2 and 5 are missing both exons 5 and 6. It has been reported by Jeffrey and co-workers<sup>28</sup> that alternative splicing produces a soluble form of HFE found predominantly in the duodenum, spleen, breast, skin and testicle and it is suggested that this soluble form may have a unique function to regulate cellular iron transport. The lack of exons five and six indicate that isoforms 2 and 5 are potentially the soluble form as they lack the transmembrane and cytoplasmic anchor. The full-length HFE was found as the predominant isoform present in the duodenum associated with hereditary hemochromatosis (HH). HH, along with porphyria cutanea tarda (PCT) and variegate porphyria (VP) are caused by a defect in HFE. It seems probable that these defects are a result of either a mistake in splicing leading to an inactive variant or as an

incorrect variant being expressed. Another factor involved could be the changes these splicing events make to the number of glycosylation sites probably affecting the trafficking of the protein through the cell. This is also linked to whether the loss of glycosylation sites affects the maturation process of the protein in such a way that a mature form is not produced. The splicing probably also affects the turnover of the proteins produced; though little is known about how exactly splicing affects this. This must be considered for all isoform models produced.

### CONCLUSION

It has been demonstrated here that, given at least one resolved structure that has homology around the splice site, it is possible to effectively model alternative splice forms. Deletions produce superior model validation values, though this may be an effect of the modelling procedure. Additions to sequences where there is little homology are more difficult to model, though not impossible, increasing in difficulty with increasing sequence length. Many of the splicing events are associated with adapting post-translational modification either in the N-terminal region by changing the signal peptide or by affecting the number or availability of glycosylation sites. Often the splicing is associated with loss or gain of whole structural units, as opposed to just changing small loop regions. The visualization of the possible structures of biomedically relevant proteins and their variants such as BACE, IL4, Frataxin, and HLF, support existing information and providing new insight into their function.

In this study mRNA and protein structures have been linked to provide a greater understanding of alternative splice forms. As the depth of transcript coverage in EST and mRNA data increases and more structures are determined, the opportunity to effectively model alternative splice forms by these methods will increase. These methods can be used to study alternative splice forms in any species with extensive transcript coverage and homologues in the PDB. Information from studies of biomedically relevant proteins could be used for tissue-specific drug development and improved diagnostics.

### ACKNOWLEDGMENTS

Financial support was provided by the United Kingdom Engineering and Physical Sciences Research Council.

### REFERENCES

- Graveley BR. Alternative splicing: increasing diversity in the proteomic world. *Trends Genet* 2001;17:100–107.
- Grabowski PJ, Black DL. Alternative splicing in the nervous system. *Prog Neurobiol* 2001;65:289–308.
- Thackeray JR, Ganetzky B. Conserved alternative splicing patterns and splicing signals in the *Drosophila* sodium channel gene *para*. *Genetics* 1995;141:203–214.
- Fettiplace R, Fuchs PA. Mechanisms of hair cell tuning. *Annu Rev Physiol* 1999;61:809–834.
- Murphy T, Yip A, Brayne C, Easton D, Evans JG, Xuereb J, Cairns N, Esiri MM, Rubinsztein DC. The BACE gene: genomic structure and candidate gene study in the late-onset Alzheimer's disease. *Neuroreport* 2001;12:631–634.
- Seah GT, Gao PS, Hopkin JM, Rook GAW. Interleukin-4 and its alternatively spliced variant (IL-4delta2) in patients with atopic asthma. *Am J Respir Crit Care Med* 2001;164:1016–1018.
- Patel PI, Isaya G. Friedreich ataxia: from GAA triplet-repeat expansion to frataxin deficiency. *Am J Hum Genet* 2001;69:15–24.
- Roy CN, Enns CA. Iron homeostasis: new tales from the crypt. *Blood* 2000;96:4020–4027.
- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucl Acids Res* 2000;28:235–242.
- Bairoch A, Apweiler R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucl Acids Res* 1999;27:49–54.
- Etzold T, Ulyanov A, Argos P. SRS: information retrieval system for molecular biology data banks. *Method Ezymol* 1996;226:114–128.
- Kersey P, Hermjakob H, Apweiler R. VARSPLIC: alternatively-spliced protein sequences derived from SWISS-PROT and TrEMBL. *Bioinformatics* 2000;16:1048–1049.
- InterBase, Version 6.0. California: Borderland Software Corp.; 2002.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990;215:403–410.
- Sali A, Blundell TL. Comparative protein modeling by satisfaction of spatial restraints. *J. Mol Biol* 1993;234:779–815.
- Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucl Acids Res* 1994;22:4673–4680.
- Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. *J Appl Crystallgr* 1993;26:283–291.
- Hutchinson EG, Thornton JM. PROMOTIF: a program to identify and analyse structural motifs in proteins. *Protein Sci* 1996;5:212–220.
- Luthy, R, Bowie, JU, Eisenberg, D. Assessment of protein models with three-dimensional profiles. *Nature* 1992;356:83–85.
- Kaplan W, Littlejohn TG. Swiss-PDB Viewer (Deep View). *Briefings in Bioinformatics* 2001;2:195–197.
- Tanahashi H, Tabira T. Three novel alternatively spliced isoforms of the human beta-site amyloid precursor protein cleaving enzyme (BACE) and their effect on amyloid beta-peptide production. *Neurosci Lett* 2001;307:9–12.
- Sauder JM, Arthur JW, Dunbrack RL. Modeling of substrate specificity of the Alzheimer's disease amyloid precursor protein  $\beta$ -secretase. *J Mol Biol* 2000;300:241–248.
- Atamas SP, Choi J, Yurovsky VV, White B. An alternative splice variant of human IL-4, IL-4 $\beta$ 2, inhibits IL-4-stimulated T cell proliferation. *J Immunol* 1996;156:435–441.
- Zav'yalov VP, Denesyuk AI, White B, Yurovsky VV, Atamas SP, Korpela T. Molecular model of an alternative splice variant of human IL-4, IL-4 $\beta$ 2, a naturally occurring inhibitor of IL-4-stimulated T cell proliferation. *Immunol Lett* 1997;58:149–152.
- Pandolfo M. Frataxin deficiency and mitochondrial dysfunction. *Mitochondrion*. Forthcoming.
- Campuzano V, Montermini L, Molto MD, Pianese L, Cossee M, Cavalcanti F, Monros E, Rodius F, Duclos F, Monticelli A, Pandolfo M. Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. *Science* 1996;271:1423–1427.
- Pianese L, Tammaro A, Turano M, De Biase I, Monticelli A, Cocza S. Identification of a novel transcript of X25, the human gene involved in Friedreich ataxia. *Neurosci Lett* 2002;320:137–140.
- Jeffrey GP, Basclain K, Hajek J, Chakrabarti S, Adams PC. Alternate splicing produces a soluble form of the hereditary hemochromatosis protein Hfe. *Blood Cell Mol Dis* 1999;25:61–67.

## APPENDIX A

**TABLE A. A Summary of the Modeling Dataset with the PDB Template Files Used and the Validation of the Modeled Splice Variants Using PROCHECK Parameters Indicating the Percentage of Residues in Core, Allowed, Generously Allowed, and Disallowed Regions Respectively Listed of the Ramachandran Plot**

Protein	SwissProt accession	Resolved structure	Ramachandran plot	Isoform models	Ramachandran plot data (%)		
Deoxyhypusine synthase	P49366	1DHS	93.6	Short isoform	90.4		
			6.0		9.3		
			0.3		0.4		
			0.0		0.0		
Frataxin	Q16595	1DLX_A	68.6	Short isoform	86.3		
			27.6		12.3		
			1.9		0.0		
			1.9		1.4		
Beta-glucuronidase	P08236	1BHG_A	71.4	Short isoform	79.1		
			23.0		15.6		
			3.7		3.4		
			1.9		1.9		
Growth factor receptor-bound protein 2	P29367	1BMB_A	89.5	Isoform GRB3-3	87.6		
			10.5		11.8		
			0.0		0.7		
			0.0		0.0		
			1CJ1_A		85.9	Isoform ASH-M	87.0
					14.1		12.4
					0.0		0.0
		1GRI	69.1		0.6		
			30.4		0.0		
			0.0		0.0		
			0.6		0.0		
		Interleukin-4	P05112	1HIK	88.6	Short isoform	88.8
					9.8		10.3
					0.8		0.9
0.8	0.0						
0.8	0.0						
1RCB	82.9				15.4		
	15.4				0.0		
	1.6				0.0		
	91.1				0.8		
	7.3				0.8		
Hereditary hemochromatosis protein	Q30201	1A6Z_A	85.4	Isoform 2	94.3		
			14.2		5.7		
			0.4		0.0		
			0.0		0.0		
			0.0		0.0		
		1DE4_G	81.2	Isoform 3	89.0		
			15.9		8.8		
			1.7		1.8		
			1.3		0.0		
			1.3		0.0		
Isoform 4			88.3		88.3		
			9.7		9.7		
			1.4		1.4		
			0.7		0.7		
Integrin beta-1 binding protein 1	O14713	1K11_A	84.8	Isoform 2	89.5		
			10.7		6.6		
			3.6		1.3		
			0.9		2.6		
			0.9		2.6		

**APPENDIX A**  
**TABLE A. (Continued)**

Protein	SwissProt accession	Resolved structure	Ramachandran plot	Isoform models	Ramachandran plot data (%)		
SH2 domain protein 1A	O60880	1D1Z_A	95.3	Isoform B	95.4		
			4.7		4.6		
			0.0		0.0		
		1D4T_A	0.0	Isoform C	90.9	89.6	
			9.1		10.4		
			0.0		0.0		
		1D4W_A	0.0	Isoform E	93.2	90.9	
			6.8		7.6		
			0.0		1.5		
			Isoform F	0.0	90.7		
					9.3		
					0.0		
		Mitogen-activated protein kinase 14	Q16539	1A9U	88.3	Isoform CSBP1	86.5
					10.7		11.0
					1.0		2.2
1BL7_A	0.0			Isoform MX12	89.8	87.6	
	9.8				10.5		
	0.3				1.5		
1DI9_A	0.0				85.1	0.4	
	14.9						
	0.0						
1P38_A	0.0				90.9		
	8.1						
	1.0						
1WFC	0.0				90.2		
	8.9						
	0.7						
Protein phosphatase 2C alpha isoform	P35813	1A6Q	90.4	Isoform Alpha-2	92.9		
			8.0		6.8		
			1.3		0.0		
Ras-related C3 botulinum toxin substrate 1	P15154	1E96_A	90.9	Isoform B	90.5		
			8.4		7.7		
			0.0		1.8		
		1FOE_F	0.6		88.9	0.0	
			11.1				
			0.0				
		1G4U_R	0.0		89.3		
			10.3				
			0.0				
		1HE1_D	0.0		92.8		
			7.2				
			0.0				
					0.0		

**APPENDIX A**  
**TABLE A. (Continued)**

Protein	SwissProt accession	Resolved structure	Ramachandran plot	Isoform models	Ramachandran plot data (%)
		1I4L_D	76.3		
			19.3		
			3.9		
			0.0		
Serine hydroxymethyltransferase	P34896	1CJO_A	88.7	Isoform 2	91.0
			10.7		7.5
			0.3		0.8
			0.3		0.8
		1BJA	90.4	Isoform 3	91.7
			9.1		7.2
			0.2		1.1
			0.2		0.0
		1EJL_A	82.5		
			15.3		
			1.7		
			0.5		
Somatotropin	P01241	1A22_A	92.6	Isoform 2	88.7
			7.4		8.2
			0.0		2.5
			0.0		0.6
		1BP3_A	75.0	Isoform 3	91.3
			22.6		4.3
			2.4		4.3
			0.0		0.0
		1HGU	60.7	Isoform 4	93.2
			33.3		4.5
			5.4		1.5
			0.6		0.8
		1HUW	95.9		
			3.4		
			0.7		
			0.0		
		1HWG_A	90.9		
			7.9		
			0.6		
			0.6		
Beta-secretase	P56817	1FKN_A	88.2	Isoform B	90.0
			11.2		8.1
			0.6		1.6
			0.0		3.0
				Isoform C	91.8
					6.8
					1.4
					0.0
				Isoform D	87.8
					10.0
					1.5
					0.7