BMC Bioinformatics

Methodology article

Identification and characterization of subfamily-specific signatures in a large protein superfamily by a hidden Markov model approach

Kevin Truong and Mitsuhiko Ikura*

Address: Division of Molecular and Structural Biology, Ontario Cancer Institute and Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada

E-mail: Kevin Truong - ktruong@uhnres.utoronto.ca; Mitsuhiko Ikura* - mikura@uhnres.utoronto.ca

*Corresponding author

Published: 10 January 2002

BMC Bioinformatics 2002, **3**:1

Received: 17 August 2001
Accepted: 10 January 2002

This article is available from: http://www.biomedcentral.com/1471-2105/3/1

Abstract

Background: Most profile and motif databases strive to classify protein sequences into a broad spectrum of protein families. The next step of such database studies should include the development of classification systems capable of distinguishing between subfamilies within a structurally and functionally diverse superfamily. This would be helpful in elucidating sequence-structure-function relationships of proteins.

Results: Here, we present a method to diagnose sequences into subfamilies by employing hidden Markov models (HMMs) to find windows of residues that are distinct among subfamilies (called signatures). The method starts with a multiple sequence alignment (MSA) of the subfamily. Then, we build a HMM database representing all sliding windows of the MSA of a fixed size. Finally, we construct a HMM histogram of the matches of each sliding window in the entire superfamily. To illustrate the efficacy of the method, we have applied the analysis to find subfamily signatures in two well-studied superfamilies: the cadherin and the EF-hand protein superfamilies. As a corollary, the HMM histograms of the analyzed subfamilies revealed information about their Ca^{2+} binding sites and loops.

Conclusions: The method is used to create HMM databases to diagnose subfamilies of protein superfamilies that complement broad profile and motif databases such as BLOCKS, PROSITE, Pfam, SMART, PRINTS and InterPro.

Background

The biological function of a protein can often be inferred from its similarity to sequences of known function in sequence databases using single-sequence similarity algorithms such as BLAST [1] or FASTA [2]. Such algorithms are suitable for determining highly similar sequences, but are not sensitive enough to capture highly divergent sequences. Therefore, many members of an evolutionarily diverse family of proteins may be overlooked. Within the last decade, the sensitivity of sequence searching techniques has been improved by profile- or motif-based analysis, which uses information derived from MSAs to construct and search for sequence patterns [3–6]. Unlike single-sequence similarity, a profile or motif can exploit additional information, such as the position and identity of residues that are conserved throughout the family, as well as variable insertion and deletion probabilities.

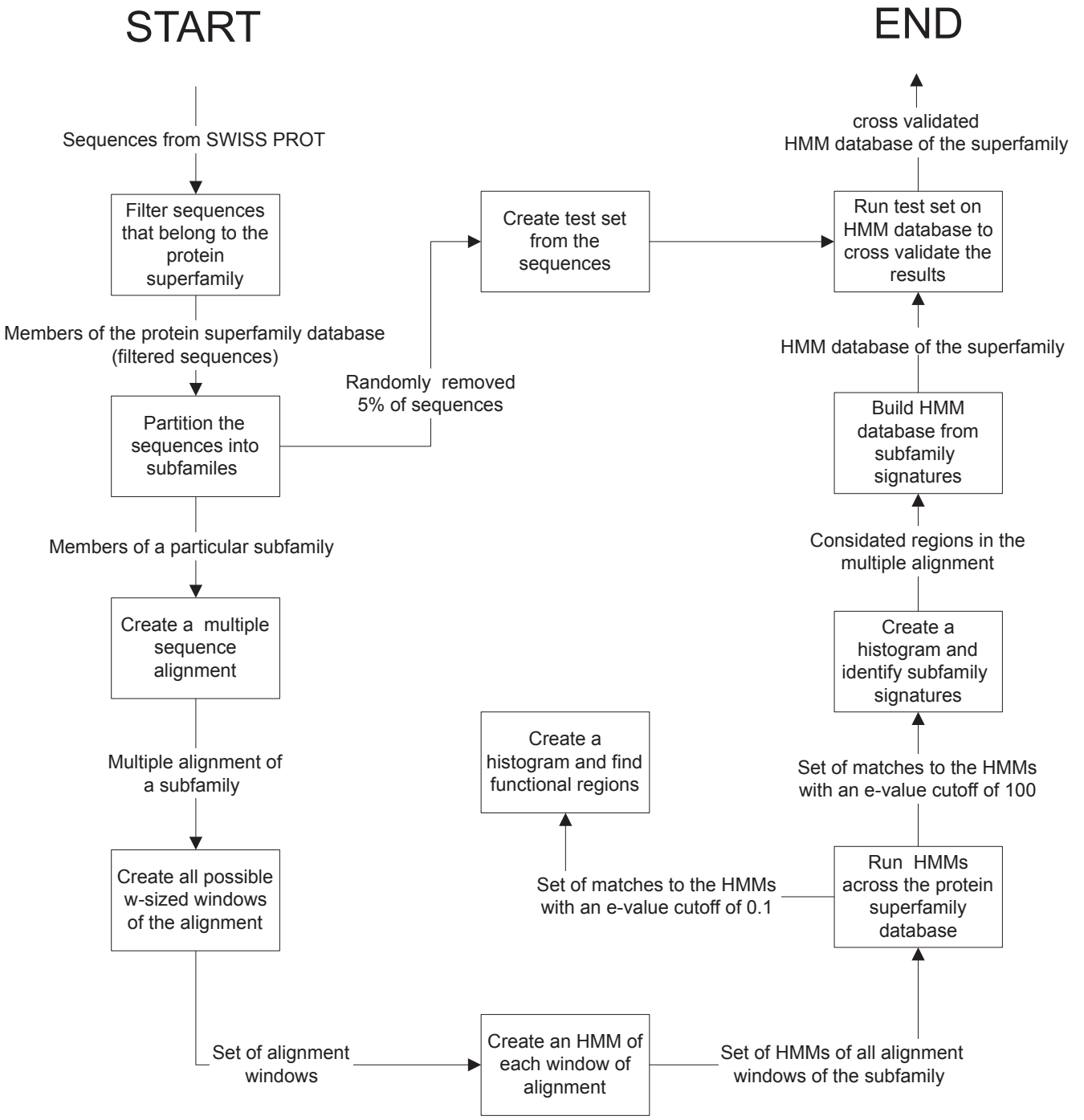


Figure 1
Flow diagram of the method. First, filter a primary database using a profile or motif database for a subset of sequences that will comprise the protein superfamily database. Then, partition the protein superfamily database into subfamilies depending on the criterion for a subfamily. Then, build an MSA for each subfamily and build HMMs of all *w* width windows of the MSA. Finally, tabulate matches with an e-value under 100 to identify subfamily signatures for the HMM database of the superfamily and tabulate matches with e-value under 0.1 to identify potentially significant functional regions in the subfamily.

Currently, the most widely-used profile and motif databases are: BLOCKS [4], which stores ungapped MSAs corresponding to the most conserved regions of protein families; PROSITE [3], which uses single consensus patterns and profiles to characterize each family of sequences; Pfam [7] or SMART [8], which uses profile hidden Markov models (HMMs) to find commonly occurring protein domains; and PRINTS-S [9], which is a database similar to both PROSITE and BLOCKS, except it uses "fingerprints" composed of more than one pattern to characterize a protein family. Recently, a new profile and motif database, InterPro [10,11], consisting of an amalgamation of PROSITE, ProDom [12], Pfam and the PRINTS fingerprint database, was used in the automatic annotation of complete proteomes including fly[13] and human[14].

Most of these databases strive to classify protein sequences into broad families, with the exception of the PRINTS-S fingerprint database, which has both family- and subfamily-specific fingerprints [9]. The ability to classify query proteins into subfamilies within superfamilies is useful in providing more specific functional annotations. Therefore, we propose a method based on HMMs to find windows of residues that are distinct in protein subfamilies. Although HMMs are expensive, both in terms of memory and computation time, they provide a solid statistical foundation for the modeling of information in an MSA. Our method works by constructing an HMM database representing a sliding window of residues for the MSA of each subfamily and then comparing the HMM database across an entire sequence database of the protein superfamily (Fig. 1). To demonstrate the utility of our approach, it has been applied to two well-studied protein superfamilies: the cadherin superfamily [15] and the EF-hand superfamily [16].

Results and discussion

Subfamily partitioning

The purpose of subfamily partitioning is to create an MSA of each subfamily, however, if quality MSAs of subfamilies already exist, it is possible to commence with the analysis at that point, as is done for PRINTS-S [9]. This section outlines a simple procedure for partitioning, however other methods exist which may be more preferable [17–21]. Many methods, like the one described herein, use a tree clustering approach based on sequence distance or identity.

The members of a protein family can be identified by collecting the matching sequences to profile or motif databases such as the ones described in the Background. This initial set of sequences is designated as the superfamily database and let the total number of sequences in this database be represented by n_T. The method of selecting a protein subfamily and defining its limits depends on the researcher who defines it. Subfamilies can be partitioned based on sequence or function and while function-based methods are valid, sequence-based methods can be automated.

To divide the sequences into subfamilies, construct a square similarity matrix, S, of dimensions n_T by n_T. $S_{i,j}$ is the percent similarity between the sequence i and sequence j. The alignment between a pair of sequences is determined in CLUSTALW by performing a global alignment [22] with an opening gap penalty of 10, an extension gap penalty of 0.1 and a Gonnet scoring matrix [23,24]. The percent similarity is estimated by the division of the alignment score by the maximum alignment score between each sequence aligned to itself.

$$S_{i,j} = \frac{alignment\_score(i,j)}{\max\left(alignment\_score(i,i), alignment\_score(j,j)\right)}$$

The similarity matrix is used to build a tree by the UPGMA (unweighted pair group method using arithmetic averages) clustering algorithm [25] for the purpose of partitioning sequences based on sequence similarity. At this point, Sjolander [20] pointed out that any partition of the tree may be meaningful. Indeed, there is no partitioning criterion that is impartially better than another. In the end, the biologist must decide the most appropriate partitioning criterion from their perspective given their experience with the protein superfamily. Therefore, the introduction of complementary methods may be important for consistent and reproducible analysis.

Our aim is to achieve a high quality MSA of each subfamily. A benchmark of the quality of an MSA is how well it reflects the structural alignment. Comparative homology modeling allows us to predict the three-dimensional structure of a target protein based on its alignment to one or more proteins with a known template structure [26]. It has been observed that as the sequence identity between the target sequence and the template increases, the average structural similarity between the template and the target also increases and for closely related protein sequences with identity over 40%, the alignment is almost always correct [27]. Therefore, if a similarity threshold greater than 40% is used for partitioning, the resulting MSAs should be reasonably high quality and well correlated with the structure. Since Dayhoff used a 60% identity for the threshold for a subfamily [17], we adopt a 60% universal similarity threshold as a slight modification. This strict threshold may create multiple partitions of the same subfamily, however, careful inspection of the sequence descriptions hint at what partitions can be joined.

Let n_S be the number of subfamilies and n_i be the number of sequences in the ith subfamily. Therefore, the number of sequences that cannot be partitioned, n_H, can be expressed in the following equation:

$$n_T = \sum_{i=1}^{n_S} n_i + n_H$$

These sequences are less than 60% similar to each other and to sequences in any subfamily. Note that n_H will never be zero due to the intermediate nodes of the initial tree. Also note that n_H will increase as the similarity of the sequences in the superfamily decreases.

Creating an HMM histogram for one subfamily

The creation of an HMM histogram for a subfamily commences with an MSA, which can be acquired from manual or automatic sequence alignment of the sequences in each subfamily. If another method was used for partitioning subfamilies, it is necessary to check if the automatically generated MSAs are correct; however, using the outlined partitioning procedure, an automatic MSA method such as CLUSTALW should produce a structurally correlated MSA, since the sequences in the subfamilies have a greater than 40% sequence identity.

Sliding MSA windows with a width of w are created. Let a_i be the width of the MSA of the ith subfamily, then the number of MSA windows for the ith subfamily, b_i, is:

$$b_i = a_i - w - 1$$

An HMM is created for each sliding MSA window of the subfamily by the HMMER software package [28]. The HMM database of the subfamily is created from the concatenation of all these individual HMMs and calibrated with a sample size of 10000 sequences. The superfamily sequence database is then searched with the HMM database and an HMM histogram is constructed from the number of matches of each window. Let the HMM histogram of the ith subfamily be represented by, $f_i(x)$, where x is the starting position of the window.

The window width (w) is a critical parameter in the generation of the HMM histogram. A small value w is desirable because it allows the features of an HMM histogram to be more evident. As the size of w increases from 20 to 80, it has the effect of smoothing the shape of the HMM histogram (Fig. 2). Empirically, it was determined that a good value of w is approximately 20 because lower values may create models that are statistically insignificant. If necessary, we suggest gradually increasing that number to achieve an acceptable balance between significance and window resolution.

Using HMM histograms to find subfamily signatures

Finding signatures involves discovering MSA windows that can distinguish this subfamily from all other subfamilies. A particular MSA window can fall into one of three categories: divergent window (a window that is not shared by the subfamily), superfamily window (shared by the superfamily), or subfamily window (shared by the subfamily). Divergent windows can be easily identified from an MSA by a stretch of positions that do not align well; however, superfamily and subfamily windows cannot be separated because they will both align well.

However, from an HMM histogram, subfamily windows have an equal number of matches ($f_i(x)$) to the number of sequences in the subfamily MSA (n_i), $f_i(x) = n_i$; superfamily windows, $f_i(x) > n_i$; divergent windows, $f_i(x) < n_i$. Since the HMM histogram sweeps across the MSA with a window size of w, if there is a subfamily signature greater than w positions, it will be identified by consecutive subfamily windows.

To define an HMM match, HMMER returns both a score and an e-value. The score is the base two logarithm of the ratio between the probability that the query sequence is a significant match to the probability that it is generated by a random model. The e-value represents the expected number of sequences with a score greater than or equal to the returned HMM score. While decreasing the e-value threshold favors finding true positives, increasing the e-value threshold favors finding true negatives. For finding subfamily signatures, a tolerant e-value of 100 is used because windows matching only sequences in the subfamily, under loose conditions, are characteristic to the subfamily.

The complete set of HMMs created from all subfamily signatures is concatenated to build the HMM database for the protein superfamily. The analysis of a query sequence follows a two-step process. First, search the query sequence for the conserved domain of the protein superfamily (i.e. presence of the cadherin repeat or EF-hand motif). If the conserved domain is found, then search for subfamily signatures. If subfamily signatures are found, the sequence belongs to the subfamily whose signature has the lowest e-value (Fig. 3). Otherwise, the sequence is classified to the protein superfamily and the classification system has achieved an equivalent level of success as most profile and motif databases. To cross-validate the analysis, remove 5% of the sequences in the initial superfamily database (the test set) prior to building the HMM histograms. The test set is checked with the constructed HMM database of the superfamily and the sequences in the test

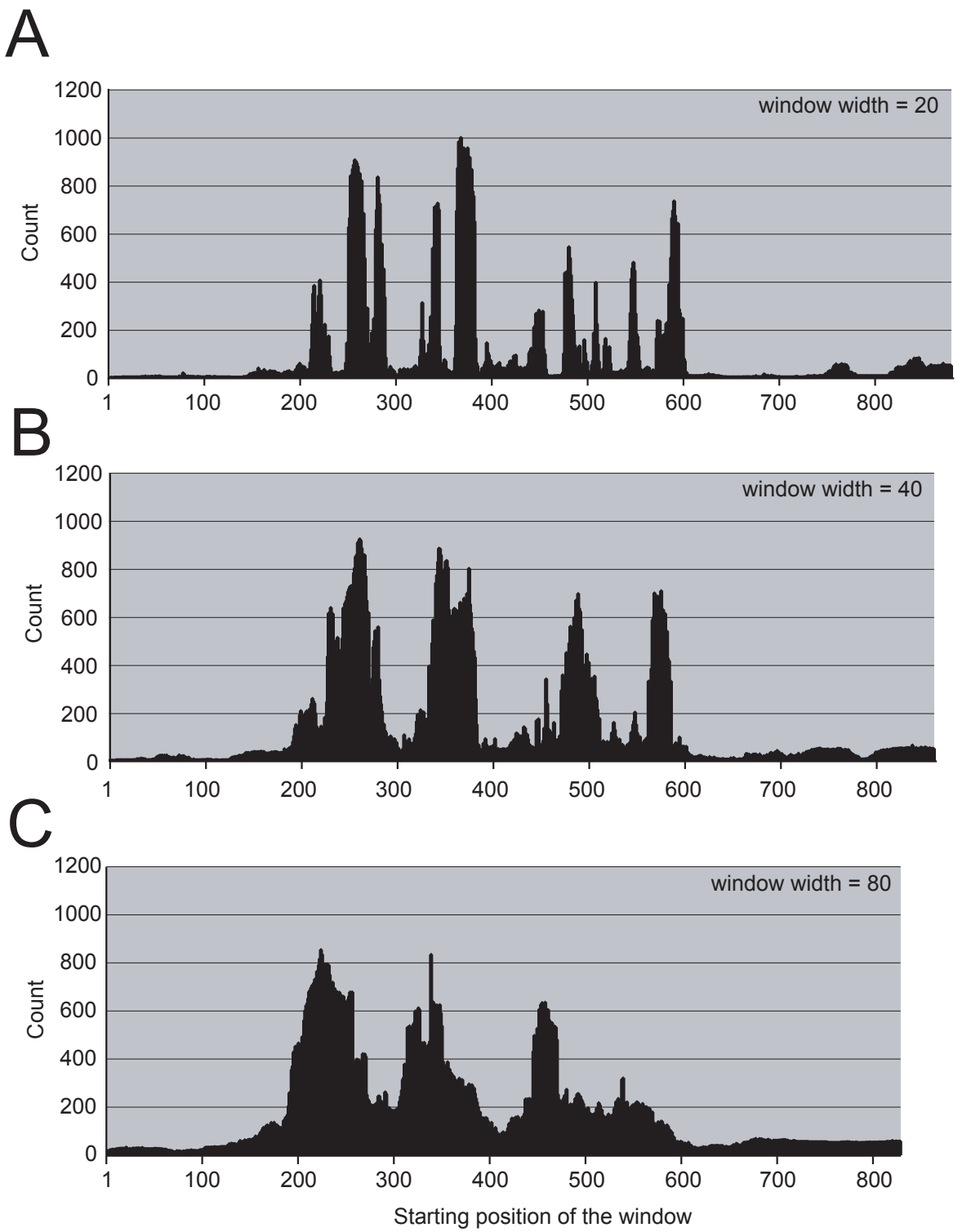


Figure 2
HMM histograms of epithelial cadherin. This figure shows the HMM histograms of epithelial cadherin with varying window widths (*w*). The x-axis represents the starting position of the window in the MSA of the subfamily; the y-axis represents the number of times that window was found in the cadherin superfamily database. The shape of the HMM histogram becomes smoother as the size of *w* increases from 20 to 40 to 80 residues because the score is calculated over a larger region.

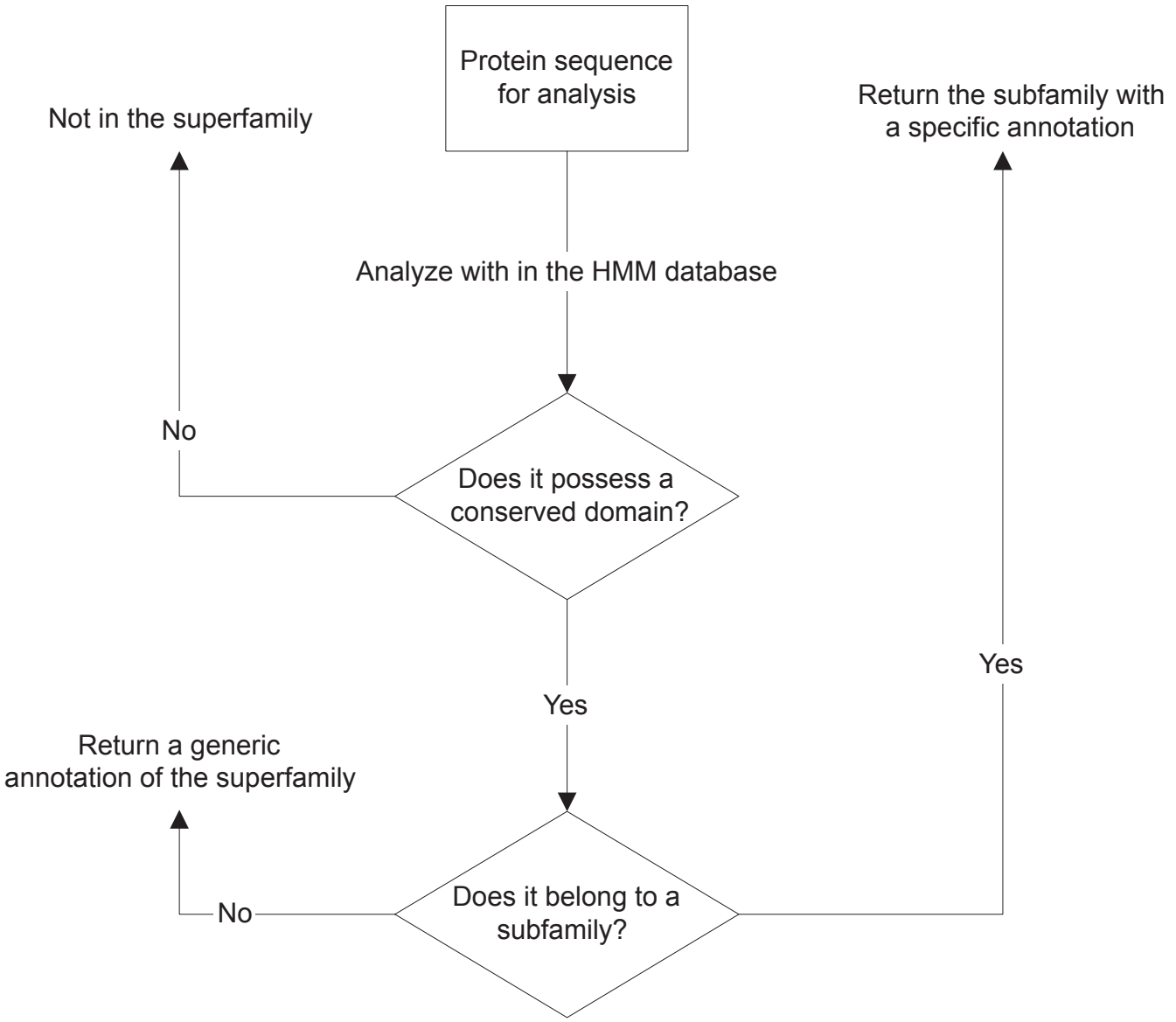


Figure 3

Flow diagram of a query into the HMM database of the superfamily

set should fall into the expected subfamilies within an acceptable error rate. We suggest a 5% acceptable error rate, but other more stringent rates may also be appropriate.

Using HMM histograms to visualize functional regions

In the previous section, to identify subfamily signatures, we focused on subfamily windows. However, superfamily windows also may provide insight into which regions in the subfamily share functional significance relative to the superfamily. Peaks in the HMM histogram can suggest which regions are particularly well conserved across the entire superfamily.

To extract this data, a few modifications are needed to the method. First, create a HMM histogram of the i^{th} subfamily as previously described, but instead with an e-value threshold of 0.1. This is a stringent threshold because for this purpose, it is important to favor true positives. Thus far, the HMM histograms presented are functions of the starting position of the window ($f_i(x)$) and while this is convenient for identifying subfamily signatures, HMM histograms as a function of the position in the alignment, $g_i(x)$, are useful to assess the contribution of individual positions.

The mapping from $f_i(x) \rightarrow g_i(x)$ is determined by tabulating a count of 1 for each position in the window when a match is found. Therefore, the mapping equation is expressed as follows:

$$g_i(x) = \sum_{n=x-w-1}^{x} f_i(n)$$

Peaks in $g_i(x)$ may hint at positions that may have functional importance.

Analysis of the cadherin superfamily

Cadherins represent a large family of proteins having diverse functions including cell-cell adhesion, morphogenesis, synapse formation, cell polarization, cell sorting, cell migration, and cell rearrangements [15]. All members of the cadherin superfamily possess a cadherin repeat (CR) and by using Pfam's HMM of the CR, 203 sequences were filtered that match the model below a 0.1 e-value from the SWISS PROT sequence database (Release 39).

Subfamily clustering produced 21 known subfamilies of cadherins with on average 8 members (Table 1). To cross-validate the effectiveness of the final HMM database of the superfamily in classifying subfamilies, 11 sequences (representing 5% of the sequence data) were removed to form the test set. The analysis to create the HMM database of the superfamily was performed using the sequences in the superfamily database minus the test set. HMM histograms of the subfamilies were created from MSAs generated by CLUSTALW (Fig. 4). 95 total subfamily signatures were extracted from the consolidation of consecutive subfamily windows. Finally, the HMM database of the superfamily was created from the concatenation of HMMs constructed from the subfamily signatures. Cross-validation revealed that all the sequences in the test set were classified into the expected subfamily.

Table 1: Tabulation of sequences in cadherin subfamilies

Cadherin subfamilies	Number of sequences (n_i)
Arcadlin	3
Desmosomal Cadherin	10
Epithelial Cadherin	7
FAT-like Cadherin	3
Flamingo Cadherin	3
Kidney Cadherin	7
Kidney Specific Cadherin	2
Liver Intestine Cadherin	4
Muscle Cadherin	2
Neural Cadherin	12
Osteoblast Cadherin	5
PB Cadherin	2
Placental Cadherin	2
Protocadherinα	28
Protocadherinβ	14
ProtocadherinγA	26
ProtocadherinγB	10
ProtocadherinγC	8
Truncated Cadherin	4
Tyrosine Receptor Kinase	6
Vascular Endothelial Cadherin	5
Unpartitioned (n_H)	**40**
Average	**8**
Total (n_T)	**203**

Table 2: Tabulation of sequences in EF-hand subfamilies

EF-hand subfamilies	Number of sequences (n_i)
Aequorin	2
α-spectrin	7
Calbindin D28k	5
Ca^{2+} Dependent Protein Kinase	33
Calcineurin B	9
Muscle Calexcitin	3
Neural Calexcitin	4
Calpain	29
Calretinin	4
Caltractin	25
Diacylglglyerol Kinase	7
α-actinin	13
Calmodulin	49
Fimbrin	3
Glycerol-3-Phosphate Dehydrogenase	4
Guanylyl Cyclase Activating Protein	5
Myosin Essential	26
Myosin Regulatory	47
P22	2
ParvalbuminA	14
ParvalbuminB	22
Phospholipase	4
Ryanodine Receptor	5
Recoverin	25
Sorcin	3
Troponin C	39
Unpartitioned (n_H)	**347**
Average	**15**
Total (n_T)	**736**

From the solved crystal structure of the first and second N-terminal CRs (CR1 and CR2) of epithelial cadherin [29], it was shown that the homodimerization of epithelial cadherin is stabilized by the Ca^{2+} ions bound in the linker region between CRs. Single amino acid substitutions in the Ca^{2+} binding site could disrupt the cell adhesion function

[30]. The HMM histogram of the epithelial cadherin subfamily was plotted on the solved crystal structure (Fig. 5A) where interestingly, the Ca^{2+} binding linker between CR1 and CR2 had the highest counts. Furthermore, the peaks of the HMM histogram were found within one or two positions in 6 of 8 residues critical in Ca^{2+} binding (Fig. 5B,C).

Various biochemical and structural studies have suggested that Ca^{2+} binding occurs between all CRs [31]. These Ca^{2+} binding linkers seem to play critical roles in the cell-adhesion function of cadherins, as they are directly involved in molecular assembly [29]. The high peak between linker of CR2 and CR3 in the HMM histogram (Fig. 5B) strongly suggests the functional importance of this domain linker. Interestingly, the two linkers between the last 3 CRs do not display an intense peak in the HMM histogram. These findings may suggest that the two N-terminal linkers are functionally more essential than the two C-terminal linkers. Further structural and mutagenesis studies are required to test this hypothesis derived from our sequence analysis.

Analysis of the EF-hand superfamily

Kretsinger and Nockolds [32] discovered the EF-hand motif in the crystal structure of parvalbumin in 1973. The EF-hand motif has a characteristic helix-loop-helix structure, consisting of approximately 30 residues. Numerous proteins that interact with Ca^{2+} contain the EF-hand motif [33]. The most prevalent classification of the EF-hand superfamily based on domain relations has been reported previously [16].

Using Pfam's HMM of the EF-hand, 736 sequences were filtered from SWISS-PROT (Rel. 39) to comprise our EF-hand superfamily database. The subfamily partitioning methodology presented here produced 26 known EF-hand subfamilies, each consisting of approximately 15 members (Table 2). The subfamily partitioning identified a significant portion of classified EF-hand subfamilies, however not all. This is because our subfamily partitioning is based entirely on sequence similarity while previous classifications utilized not only sequence similarities but also other information available from experimental studies. In addition, there was a large portion of the superfamily which could not be partitioned using strictly sequence similarity, suggesting that sequences in the EF-hand superfamily are significantly dissimilar and that a complementary approach may be need to fully partition all subfamilies.

Similar to the cross-validation analysis on the cadherin superfamily, 37 sequences (representing 5% of the sequence data) were removed to form the test set. Again, HMM histograms of the subfamilies were created from the reduced set of superfamily sequences (Fig. 6). In total, 40 subfamily signatures were extracted. The HMM database of the EF-hand superfamily was created from the subfamily signatures. Again, cross-validation revealed that all the sequences in the test set were classified into the expected subfamily. This suggested that the method can classify sequences with a high specificity.

The peaks in the HMM histograms corresponded to windows that include EF-hand motifs (Fig. 6). Calbindin D28k, for example, has six EF-hands (designated EF1-EF6). Ca^{2+} binding studies have shown that EF2 does not bind Ca^{2+} and EF6 binds Ca^{2+} with a lower affinity than the other four functional sites [34]. Interestingly, the HMM histogram of Calbindin D28k shows no peaks at the locations of EF2 and EF6 (Fig. 6A). Calcineurin B contains four EF-hands, all shown to bind Ca^{2+}[35]. The HMM histogram clearly shows the presence of four functionally active Ca^{2+} binding EF-hands in calcineurin B (Fig. 6B). Caltractin also possesses four EF-hands: two higher affinity and two lower affinity [36]. Similarly, the HMM histogram shows the four peaks corresponding to four EF-hands (Fig. 6C). Parvalbumin is a Ca^{2+} buffering protein involved in the relaxation of muscle after contraction by binding up free Ca^{2+} in the cell [37,38]. The HMM histogram was mapped onto the solved crystal structure of parvalbumin B [39] (Fig. 7A). Parvalbumin B has three EF-hands and the first N-terminal EF-hand does not bind Ca^{2+}[39]. The HMM histogram clearly displayed the lack of the functional N-terminal EF-hand and the existence of two active C-terminal EF-hands (Fig. 7B,C). These examples demonstrated that HMM histograms are not only useful for finding subfamily signatures but also in locating functionally significant regions of subfamilies.

Conclusions

We developed a method to decipher signature regions of protein subfamilies, which can be used to build HMM databases for diagnosing subfamilies of large protein superfamilies. Using this method, we identified subfamily signatures and built HMM databases for two well-studied superfamilies of cadherins and EF-hand proteins. Additionally, peaks in the HMM histogram plots of subfamilies were found to coincide with functionally important regions (i.e. Ca^{2+} binding sites and loops). Future work should include the comparison between different subfamily partitioning techniques and also the creation of richly annotated databases for subfamilies of superfamilies for possible application in automated genomic annotation in conjunction with other motif and profile databases.

Materials and methods

The studies were performed using a variety of tools and whenever necessary, in-house programs were written to

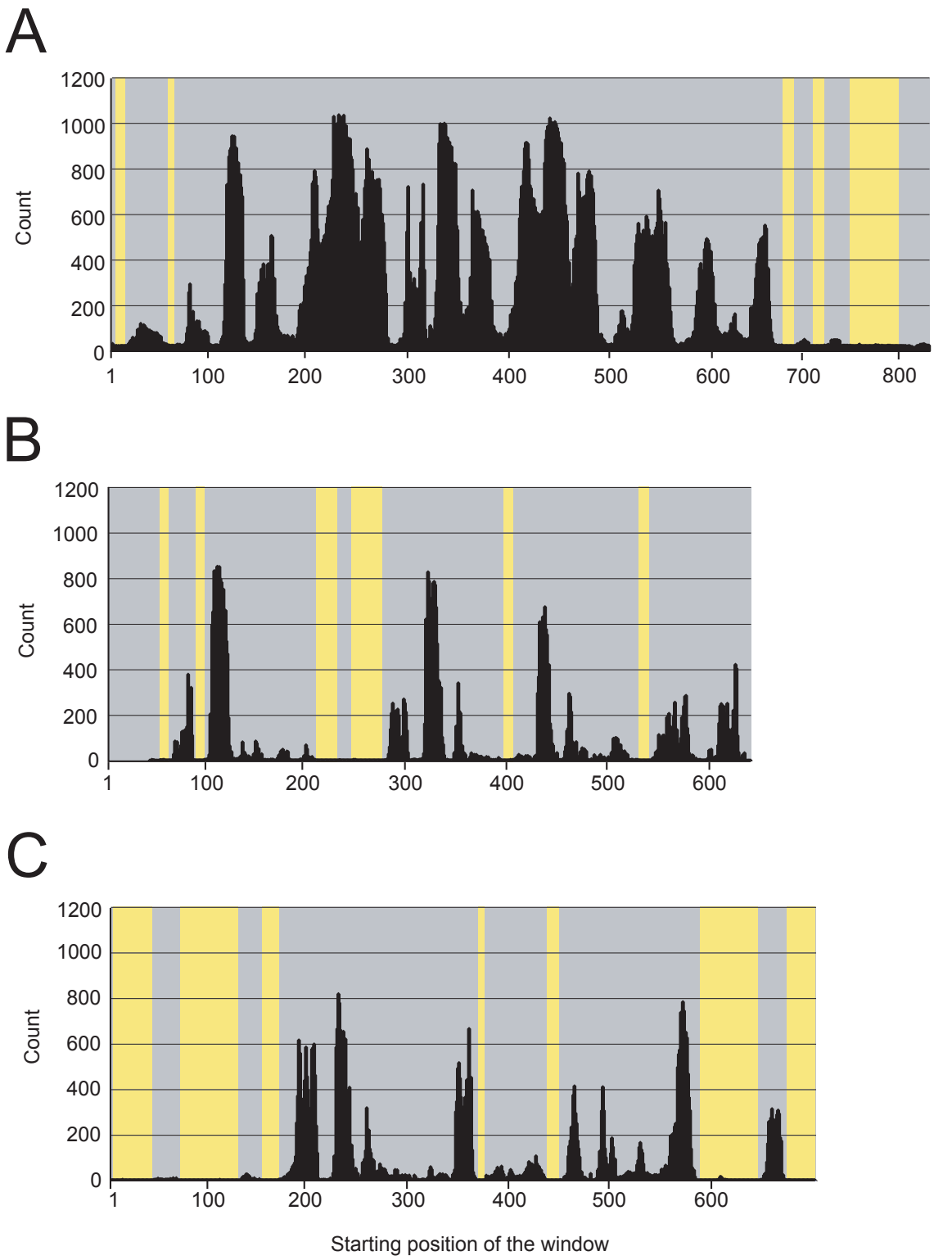


Figure 4
HMM histograms of cadherin subfamilies. The HMM histograms were constructed with a window width of 20 and an e-value threshold of 100. The signature regions are highlighted in yellow for various subfamilies in the cadherin superfamily. **A)** Proto-cadherin-γA **B)** Liver Intestine cadherin**C)** Truncated cadherin

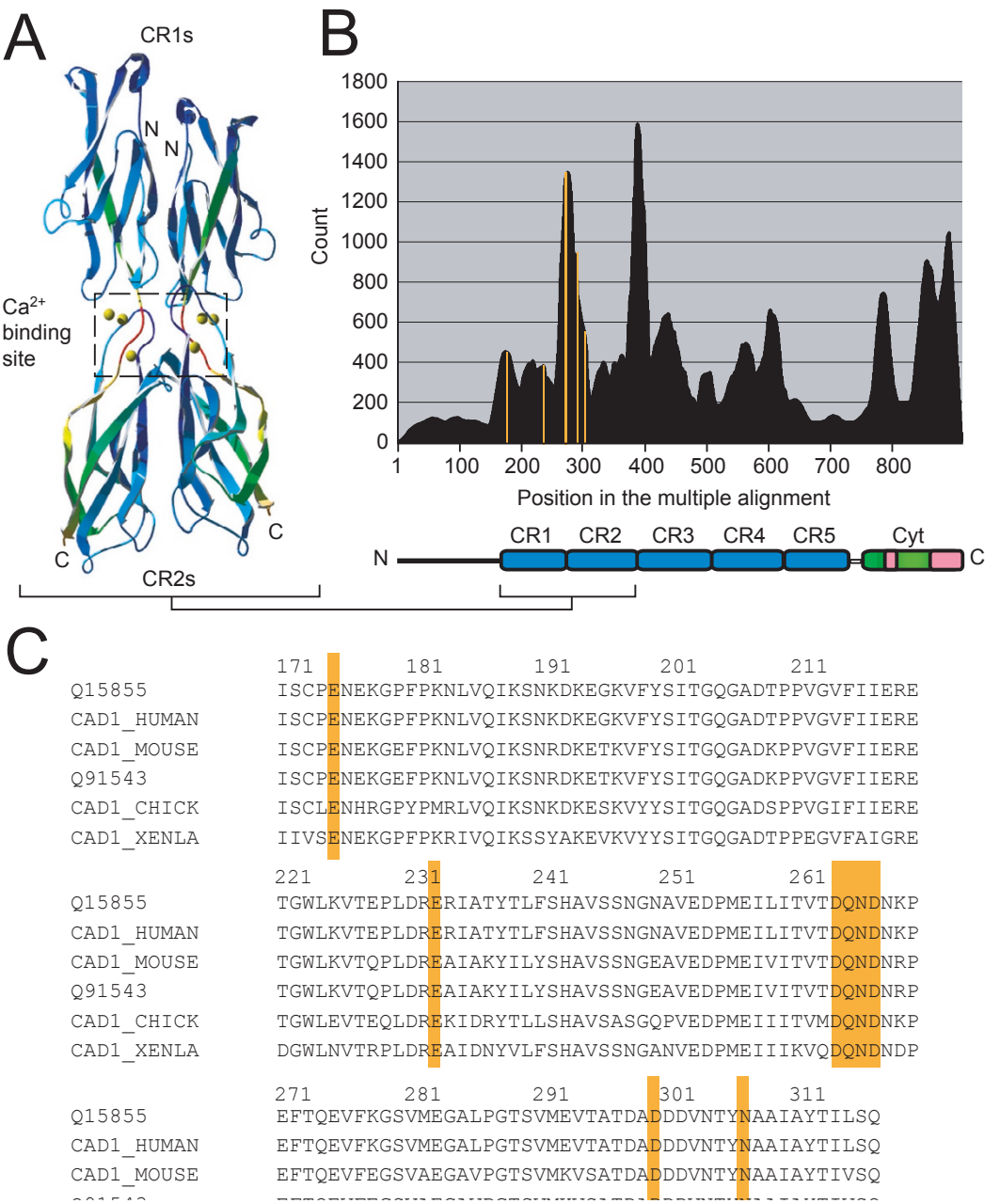


Figure 5
Mapping the HMM histogram to the crystal structure of epithelial cadherin. **A)** The HMM histogram mapped onto the crystal structure (PDB code: 1EDH) of the first and second cadherin repeats of epithelial cadherin. Ca^{2+} ions are depicted as yellow spheres. The regions of high occurrence map to the Ca^{2+} binding site (blue represents low occurrence and red represents high occurrence)**B)** The HMM histogram of the epithelial cadherin subfamily with an e-value cutoff of 0.1. The orange bars in the histogram reflect positions involved in Ca^{2+} binding. Below the histogram is the domain layout. The features are colored: cadherin repeat (CR), blue rectangle; cytoplasmic domain (Cyt), green rectangle; catenin binding sites in the cytoplasmic domain, pink rectangles. The segment between the last CR and the cytoplasmic domain is the single pass transmembrane domain.**C)** The MSA of the segment involved in Ca^{2+} binding between the first and second CRs. The SWISS PROT code of the sequence is shown in the left and the 8 residues involved in Ca^{2+} binding are highlighted orange.

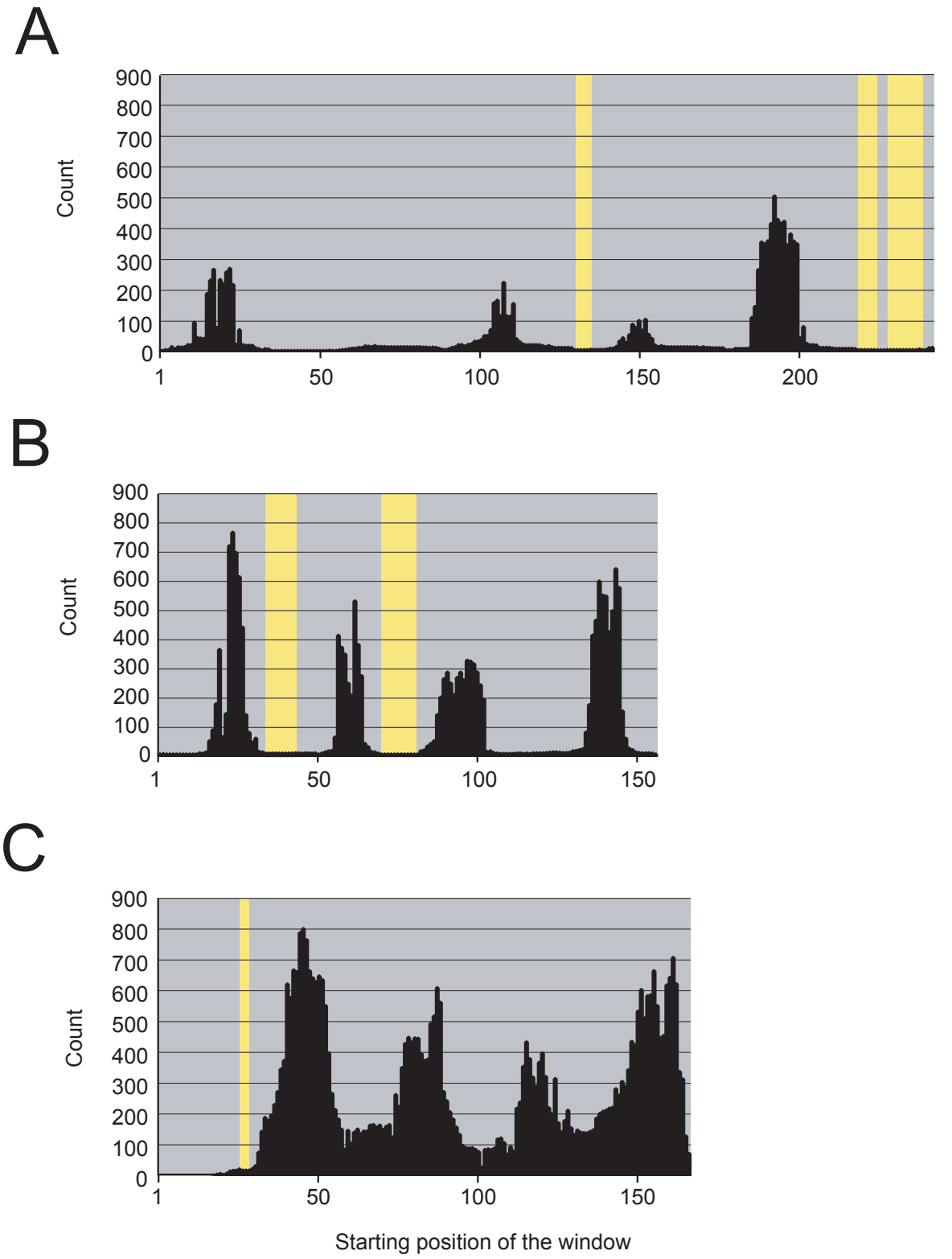
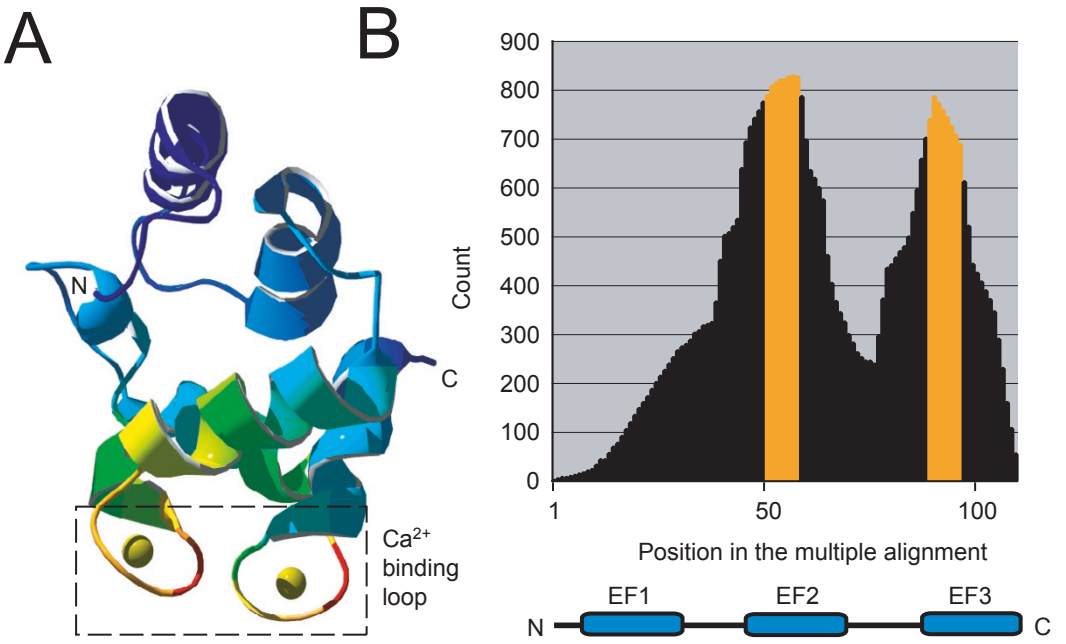


Figure 6
HMM histograms of EF-hand subfamilies. Using the same conventions as Fig. 4, the HMM histograms were constructed for various subfamilies in the EF-hand superfamily. **A)** Calbindin D28k **B)** Calcineurin B **C)** Caltractin.



C

```
                1          11         21         31         41         51
PRVB_GRAGE      AMTDILSAKDIEAALTSCQAADSFNYKSFFSKVGLKGKSTDQVKKIFGILDQDKSGFIEE
PRVB_AMPME      AITDILSAKDIEAALSSVKAAESFNYKTFFTKCGLAGKPTDQVKKVFDILDQDKSGYIEE
PRVB_RANES      SITDIVSEKDIDAALESVKAAGSFNYKIFFQKVGLAGKSAADAKKVFEILDRDKSGFIEQ
PRVB_ESOLU      SFAG-LKDADVAAALAACSAADSFKHKEFFAKVGLASKSLDDVKKAFYVIDQDKSGFIEE
PRVB_OPSTA      SFAGILSDADIDAALAACQAAESFKHKEFFAKVGLSAKTPDVIKKAFYVIDQDKSGFIEE
PRVB_MERME      AFAGILADADITAALAACKAEGSFKHGEFFTKIGLKGKSAADIKKVFGIIDQDKSDFVEE
PRVB_MERBI      AFSGILADADVAAALKACEAADSFNYKAFFAKVGLTAKSADDIKKAFFVIDQDKSGFIEE
PRVB_MERMR      AFAGILADADCAAAVKACEAADSFSYKAFFAKCGLSGKSADDIKKAFVFIDQDKSGFIEE
PRVB_CYPCA      AFAGVLNDADIAAALEACKAADSFNHKAFFAKVGLTSKSADDVKKAFAIIDQDKSGFIEE
PRVB_LEUCE      --AFGLKEADITAALEACKAADSFNHKAFFAKVGMSAKSAGDVKKAFEIIDEDKSGFIEE
PRVB_LATCH      AVAKLLAAADVTAALEGCKADDSFNHKVFFQKTGLAKKSNEELEAIFKILDQDKSGFIEE

                61         71         81         91         101
PRVB_GRAGE      DELQLFLQNFSSTARALTAAETKAFMAAGDTDGDGKIGVDEFQALVKA-
PRVB_AMPME      DELQLFLKNFCSSARSLSNAETKAFLFAGDSDGDGKIGVDEFQALVRS-
PRVB_RANES      DELGLFLQNFRASARVLSDAETSAFLKAGDSDGDGKIGVEEFQALVKA-
PRVB_ESOLU      DELKLFLQNFSPSARALTDAETKAFLADGDKDGDGMIGVDEFAAMIKA-
PRVB_OPSTA      DELKLFLQNFASSARALTDKETETFLKAGDSDGDGKIGIDEFADLVKEA
PRVB_MERME      DELKLFLQNFSAGARALTDAETATFLKAGDSDGDGKIGVEEFAAMVKG-
PRVB_MERBI      DELKLFLQVFSAGARALTDAETKAFLKAGDSDGDGAIGVDEWAALVKA-
PRVB_MERMR      DELKLFLQVFKAGARALTDAETKAFLKAGDSDGDGAIGVEEWVALVKA-
PRVB_CYPCA      DELKLFLQNFKADARALTDGETKTFLKAGDSDGDGKIGVDEFTALVKA-
PRVB_LEUCE      EELKLFLQNFKAGARALTDAETKIFLKAGDADGDGKIGIDEFAALVKA-
PRVB_LATCH      DEELELFLQNFSAGARTLTKTETETFLKAGDSDGDGKIGVDEFQKLVKA
```

Figure 7
Mapping the HMM histogram to the crystal structure of parvalbumin B. **A)** The HMM histogram mapped onto the crystal structure (PDB code: 1CDP) of parvalbumin B. Using the same conventions as Fig. 5A, the Ca^{2+} binding loops of two EF-hand motifs have a high occurrence level. **B)** The HMM histogram of the parvalbumin B subfamily with an e-value cutoff of 0.1. The orange regions in the histogram reflect the segments encoding the Ca^{2+} binding loops. Below the histogram is the domain layout. The blue rectangle represents the EF-hand motif (EF). **C)** The MSA of the parvalbumin B subfamily. Using the same conventions as Fig. 5C, the Ca^{2+} binding loops are highlighted orange.

pre- and post-process data from the different applications. MSAs were generated using CLUSTALW [40] and all HMMs were created using the HMMER package [28]. Data was stored on the Oracle relational database management system and Microsoft FoxPro was used as an ODBC (Open Database Connectivity) client for querying and joining tables from the database. Microsoft Excel was used for dynamic charting of data. Perl was used for shell scripting, text manipulation and pattern matching with regular expressions. HMMER, CLUSTALW, Oracle database server (version 8) and Perl scripts were executed on a machine with a dual 750 MHz UltraSPARC-111 processor and 4 G of RAM running SunOS 5.8. Microsoft FoxPro and Excel were executed on a 500 MHz Intel Celeron processor and 128 MB of RAM running a Windows 98 operating system.

The time required to analyze one superfamily depended largely on the computation platform, the number of sequences of the superfamily and the average width of subfamily MSAs. Using the computation platforms described, the computation time to generate the MSA using CLUSTALW for the cadherin superfamily (~200 sequences, ~800 average width) was ~3 hours and for the EF-hand superfamily (~700 sequence, ~200 average width) was ~9 hours. The computation time for the creation of a calibrated HMM database (window size of 20) for an average cadherin subfamily was ~6 hours; for an EF-hand subfamily, ~45 minutes. The execution time for an average HMM database of cadherin subfamily over the superfamily database was ~12 hours; for an EF-hand sub-family, ~7 hours. The computation time was extensive but could easily be adapted to a parallel computing system.

The HMM database created for the cadherin and EF-hand superfamilies and all glue programs that were used for the analysis are available upon request.

Acknowledgements

We would like to thank Gil Prive for critical reading of the manuscript. This work was supported by grants to MI from the National Cancer Institute of Canada. MI is a HHMI (Howard Hughes Medical Institute) International Scholar and CIHR (Canadian Institute of Health Research) Scientist.

References

1. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**:3389-3402
2. Pearson WR: **Using the FASTA program to search protein and DNA sequence databases.** *Methods Mol Biol* 1994, **25**:365-389
3. Hofmann K, Bucher P, Falquet L, Bairoch A: **The PROSITE database, its status in 1999.** *Nucleic Acids Res* 1999, **27**:215-219
4. Henikoff S, Henikoff JG, Pietrokovski S: **Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations.** *Bioinformatics* 1999, **15**:471-479
5. Barton GJ: **Protein multiple sequence alignment and flexible pattern matching.** *Methods Enzymol* 1990, **183**:403-428
6. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14**:755-763
7. Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL: **The Pfam protein families database.** *Nucleic Acids Res* 2000, **28**:263-266
8. Schultz J, Milpetz F, Bork P, Ponting CP: **SMART, a simple modular architecture research tool: identification of signaling domains.** *Proc Natl Acad Sci USA* 1998, **95**:5857-5864
9. Attwood TK, Croning MD, Flower DR, Lewis AP, Mabey JE, Scordis P, Selley JN, Wright W: **PRINTS-S: the database formerly known as PRINTS.** *Nucleic Acids Res* 2000, **28**:225-227
10. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MD, Durbin R, Falquet L, Fleischmann W, Gouzy J, Hermjakob H, Hulo N, Jonassen I, Kahn D, Kanapin A, Karavidopoulou Y, Lopez R, Marx B, Mulder NJ, Oinn TM, Pagni M, Servant F: **The InterPro database, an integrated documentation resource for protein families, domains and functional sites.** *Nucleic Acids Res* 2001, **29**:37-40
11. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MD, Durbin R, Falquet L, Fleischmann W, Gouzy J, Hermjakob H, Hulo N, Jonassen I, Kahn D, Kanapin A, Karavidopoulou Y, Lopez R, Marx B, Mulder NJ, Oinn TM, Pagni M, Servant F, Sigrist CJ, Zdobnov EM: **InterPro–an integrated documentation resource for protein families, domains and functional sites.** *Bioinformatics* 2000, **16**:1145-1150
12. Corpet F, Servant F, Gouzy J, Kahn D: **ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons.** *Nucleic Acids Res* 2000, **28**:267-269
13. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, George RA, Lewis SE, Richards S, Ashburner M, Henderson SN, Sutton GG, Wortman JR, Yandell MD, Zhang Q, Chen LX, Brandon RC, Rogers YH, Blazej RG, Champe M, Pfeiffer BD, Wan KH, Doyle C, Baxter EG, Helt G, Nelson CR, Gabor GL, Abril JF, Agbayani A, An HJ, Andrews-Pfannkoch C, Baldwin D, Ballew RM, Basu A, Baxendale J, Bayraktaroglu L, Beasley EM, Beeson KY, Benos PV, Berman BP, Bhandari D, Bolshakov S, Borkova D, Botchan MR, Bouck J, Brokstein P, Brottier P, Burtis KC, Busam DA, Butler H, Cadieu E, Center A, Chandra I, Cherry JM, Cawley S, Dahlke C, Davenport LB, Davies P, de Pablos B, Delcher A, Deng Z, Mays AD, Dew I, Dietz SM, Dodson K, Doup LE, Downes M, Dugan-Rocha S, Dunkov BC, Dunn P, Durbin KJ, Evangelista CC, Ferraz C, Ferriera S, Fleischmann W, Fosler C, Gabrielian AE, Garg NS, Gelbart WM, Glasser K, Glodek A, Gong F, Gorrell JH, Gu Z, Guan P, Harris M, Harris NL, Harvey D, Heiman TJ, Hernandez JR, Houck J, Hostin D, Houston KA, Howland TJ, Wei MH, Ibegwam C, *et al*: **The genome sequence of Drosophila melanogaster.** *Science* 2000, **287**:2185-2195
14. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, *et al*: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409**:860-921
15. Tepass U, Truong K, Godt D, Ikura M, Peifer M: **Cadherins in embryonic and neural morphogenesis.** *Nature Review* 2000, **1**:91-100
16. Kawasaki H, Nakayama S, Kretsinger RH: **Classification and evolution of EF-hand proteins.** *Biometals* 1998, **11**:277-295
17. Dayhoff MO: **The origin and evolution of protein superfamilies.** *Fed Proc* 1976, **35**:2132-2138
18. Lichtarge O, Bourne HR, Cohen FE: **An evolutionary trace method defines binding surfaces common to protein families.** *J Mol Biol* 1996, **257**:342-358

19. Corpet F, Gouzy J, Kahn D: **Browsing protein families via the 'Rich Family Description' format.** *Bioinformatics* 1999, **15**:1020-1027
20. Sjolander K: **Phylogenetic inference in protein superfamilies: analysis of SH2 domains.** *Proc Int Conf Intell Syst Mol Biol* 1998, **6**:165-174
21. Wicker N, Perrin GR, Thierry JC, Poch O: **Secator: a program for inferring protein subfamilies from phylogenetic trees.** *Mol Biol Evol* 2001, **18**:1435-1441
22. Needleman SB, Wunsch CD: **A general method applicable to the search for similarities in the amino acid sequence of two proteins.** *J Mol Biol* 1970, **48**:443-453
23. Benner SA, Cohen MA, Gonnet GH: **Amino acid substitution during functionally constrained divergent evolution of protein sequences.** *Protein Eng* 1994, **7**:1323-1332
24. Barker WC, Ketcham LK, Dayhoff MO: **A comprehensive examination of protein sequences for evidence of internal gene duplication.** *J Mol Evol* 1978, **10**:265-281
25. Prager EM, Wilson AC: **Construction of phylogenetic trees for proteins and nucleic acids: empirical evaluation of alternative matrix methods.** *J Mol Evol* 1978, **11**:129-142
26. Sanchez R, Sali A: **Comparative protein structure modeling. Introduction and practical examples with modeller.** *Methods Mol Biol* 2000, **143**:97-129
27. Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A: **Comparative protein structure modeling of genes and genomes.** *Annu Rev Biophys Biomol Struct* 2000, **29**:291-325
28. Eddy SR: **The HMMER Package.** 1995
29. Nagar B, Overduin M, Ikura M, Rini JM: **Structural basis of Ca^{2+}-induced E-cadherin rigidification and dimerization.** *Nature* 1996, **380**:360-364
30. Ozawa M, Engel J, Kemler R: **Single amino acid substitutions in one Ca^{2+}-binding site of uvomorulin abolish the adhesive function.** *Cell* 1990, **63**:1033-1038
31. Alattia JR, Kurokawa H, Ikura M: **Structural view of cadherin-mediated cell-cell adhesion.** *Cell Mol Life Sci* 1999, **55**:359-367
32. Kretsinger RH, Nockolds CE: **Carp muscle Ca^{2+}-binding protein. II. Structure determination and general description.** *J Biol Chem* 1973, **248**:3313-3326
33. Lewit-Bentley A, Rety S: **EF-hand Ca^{2+}-binding proteins.** *Curr Opin Struct* 2000, **10**:637-643
34. Akerfeldt KS, Coyne AN, Wilk RR, Thulin E, Linse S: **Ca^{2+}-binding stoichiometry of calbindin D28k as assessed by spectroscopic analyses of synthetic peptide fragments.** *Biochemistry* 1996, **35**:3662-3669
35. Kakalis LT, Kennedy M, Sikkink R, Rusnak F, Armitage IM: **Characterization of the Ca^{2+}-binding sites of calcineurin B.** *FEBS Lett* 1995, **362**:55-58
36. Weber C, Lee VD, Chazin WJ, Huang B: **High level expression in Escherichia coli and characterization of the EF-hand Ca^{2+}-binding protein caltractin.** *J Biol Chem* 1994, **269**:15795-15802
37. Lannergren J, Elzinga G, Stienen GJ: **Force relaxation, labile heat and parvalbumin content of skeletal muscle fibres of Xenopus laevis.** *J Physiol* 1993, **463**:123-140
38. Muntener M, Kaser L, Weber J, Berchtold MW: **Increase of skeletal muscle relaxation speed by direct injection of parvalbumin cDNA.** *Proc Natl Acad Sci U S A* 1995, **92**:6504-6508
39. Swain AL, Kretsinger RH, Amma EL: **Restrained least squares refinement of native (Ca^{2+}) and Cd- substituted carp parvalbumin using X-ray crystallographic data at 1.6-Å resolution.** *J Biol Chem* 1989, **264**:16620-16628
40. 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.** *Nucleic Acids Res* 1994, **22**:4673-4680