Phylogenetic inference in protein superfamilies: Analysis of SH2 domains

Kimmen Sjölander*
Molecular Applications Group
P.O. Box 51110
Palo Alto, CA 94303-1110
kimmen@mag.com

Abstract

This work focuses on the inference of evolutionary relationships in protein superfamilies, and the uses of these relationships to identify key positions in the structure, to infer attributes on the basis of evolutionary distance, and to identify potential errors in sequence annotations. Relative entropy, a distance metric from information theory, is used in combination with Dirichlet mixture priors to estimate a phylogenetic tree for a set of proteins. This method infers key structural or functional positions in the molecule, and guides the tree topology to preserve these important positions within subtrees. Minimum-description-length principles are used to determine a cut of the tree into subtrees, to identify the subfamilies in the data. This method is demonstrated on SH2-domain containing proteins, resulting in a new subfamily assignment for Src2_drome and a suggested evolutionary relationship between Nck_human and Drk_drome, Sem5_caeel, Grb2_human and Grb2_chick.

Introduction

Gene duplication events have played a major role in the evolution of the human genome (Miklos and Rubin 1996). Genes related in this way are called *paralogous*; groups of these paralogs form superfamilies of related genes. Each duplication event allows a freeing of functional constraints on one copy, so that over time and large evolutionary distances, a plethora of functions and structures can evolve from a single ancestor gene.

To the protein sequence analyst, these superfamilies contain a wealth of hidden information, and pose a multitude of questions. How did this family evolve? What was the ancestor protein like? What was its original function? Within this large superfamily are there subgroups defined by common functions or other attributes? If the proteins interact with other molecules, can we identify the residues involved in the binding pockets, and pinpoint those residues contributing to the substrate specificity of subfamilies in the data? Can we extrapolate from attributes that are known for only some members of a group to predict the attributes of other members for which less is known?

In this paper, I give an overview of a new method for phylogenetic reconstruction described in detail elsewhere (Sjölander 1998; 1997). Bayesian Evolutionary Tree Estimation (Bête) employs Bayesian and information-theoretic measures to construct a phylogenetic tree and identify subfamilies. Once we have a decomposition of a protein superfamily into subfamilies, we can use this decomposition to predict residues involved in the subfamily-specificity of protein function or structure, to infer attributes on the basis of evolutionary relationships and flag potential errors in sequence annotation.

Proofs of statistical consistency only exist for a limited range of models of evolution, and most assume that the sites evolve under identical processes (Erdos *et al.* 1997). On the other hand, performance under non-identical evolutionary processes (i.e., allowing rates across sites to vary) can still be quite good for some methods under some experimental conditions (Tateno *et al.* 1994; Felsenstein 1996; Hasegawa *et al.* 1991; Yang 1994; Kuhner and Felsenstein 1994). Although external biological information concerning site variability, when available, can be given as input to a phylogenetic reconstruction program, such information is often incomplete or not available, and it is clearly helpful to the methods' performance to determine rate variability from the primary sequence alone. The method presented here addresses this, and also a somewhat more general question: how to identify sites which are strongly conserved *within* subfamilies, even if they vary *between* different subfamilies.

This method has three underyling assumptions: (1) evolution conserves function and structure; (2) not all positions in a molecule are created equal, and some are more important than others in maintaining a protein's structure or function; (3) a tree that groups proteins together which are similar in key functional or structural positions is more likely to correspond to both the histor-

ical evolutionary processes underlying the data and to the inherent functional and structural hierarchy in the data.

Some sites in a protein, such as the catalytic triad of serine proteases, show perfect conservation over very large evolutionary distances. While such perfectly conserved positions are obviously essential for maintaining the protein function or structure, they are not particularly informative in differentiating among alternative tree topologies. By contrast, other binding-pocket residues may show a *subfamily-specific* conservation pattern: essentially conserved within each subfamily, but differing across subfamilies (Casari *et al.* 1995; Lichtarge *et al.* 1996).

The method described here has been developed specifically for protein superfamily analysis, to identify these positions from the primary sequence alone, and to weight these positions as more important when producing a tree topology for a family.

Rather than attempt a detailed comparison of this method with other methods over a large number of families (a task normally accomplished with simulated data), this paper focuses on a single family of proteins for which much is known of the structure and function of individual members: SH2 domains. As anyone who has compared different phylogenetic reconstruction methods is aware, tree topology reconstruction is sensitive to small errors in the multiple sequence alignment, to the inclusion of false homologs, and other complications of actual biological data. Apply three different methods to identical data, and you are likely to obtain three–or more!–different tree topologies. This is especially true with protein families having any significant degree of primary sequence diversity. Nevertheless, phylogenetic reconstruction can be a powerful tool in protein superfamily analysis.

Based on the limited nature of this comparison, no claims can be made for the superiority of this method over others. Differences in tree topologies across the methods are made primarily to illustrate (1) the high degree of uncertainty in phylogenetic reconstruction in protein superfamily analysis, and (2) the importance of recognizing sites involved in the subfamily specificity of protein function and structure in constraining tree topologies.

Method

The algorithm employed in this work to identify the functional subfamilies in a set of protein sequences can be decomposed into two subtasks: constructing an evolutionary tree, and cutting the tree into subtrees to infer the subfamilies.

Bayesian evolutionary tree estimation (Bête)

The method described here to construct the tree falls within a hierarchical clustering paradigm known as *agglomerative* clustering using *nearest neighbor* heuristics. Initially, each sequence is in its own class, and forms a leaf of the tree. At each iteration of the algorithm, the two closest classes are merged, until at termination all sequences are in a single class, forming the root of the tree. Two aspects of the method differentiate it from standard neighbor-joining tree algorithms: the representation of each class at each iteration of the algorithm, and the distance measure between classes used to choose which two classes to join.

Classes are represented by profiles, employing Dirichlet mixture priors (Sjölander *et al.* 1996) to compute the amino acid distributions at each position. Dirichlet mixture priors have been found to be highly effective at increasing the sensitivity and specificity of remote homolog identification (Karplus *et al.* 1997; Tatusov *et al.* 1994; Bailey and Elkan 1995; Brown *et al.* 1993). This makes them appealing for forming statistical models of groups of sequences during the agglomeration algorithm. In contrast to substitution matrices, which generalize all distributions to allow for substitutions of similar amino acids, Dirichlet mixture priors (and Bayesian methods in general) allow substitutions when few observations are available, but converge on the frequencies in the data as the number of observations increase. Because of this, during the agglomeration process, as increasingly divergent sequences are added to the classes being formed, conserved positions start to become evident in the profiles being formed.

The distance measure between profiles employed in this method is a symmetrized form of relative entropy (Cover and Thomas 1991), summed over all the columns in the multiple alignment. This metric, the Total Relative Entropy (TRE), is defined to be

$$TRE = \sum_c D(i_c \,||\, j_c) + D(j_c \,||\, i_c) \qquad (1)$$

where i_c and j_c are the probability distributions at position c in the profile for the i^{th} and j^{th} classes respectively, and the relative entropy between two distributions p and q is defined to be

$$D(p||q) = \sum_x p(x) log \frac{p(x)}{q(x)} \; . \qquad (2)$$

As Table I shows, Dirichlet mixture priors and relative entropy function together as an implicit weighting scheme on the columns in the multiple alignment to favor joining two classes if they are similar (or identical) at positions showing high conservation (low tolerance for mutation), and favor keeping two classes separate if they are dissimilar at such positions. These conserved positions have a large impact on the TRE between two profiles; positions showing higher tolerance for mutation (the more mixed distributions) have less impact on the TRE between two profiles. This helps constrain the tree topologies produced to maintain conserved distributions within subtrees corresponding to functional subfamilies,

and construct tree topologies that reflect the functional hierarchy in the data.

The algorithm The input to the program is a multiple sequence alignment. Obviously, the better the alignment, the more accurate the resulting tree topology. In the section that follows, the term "sequence" refers to a row in the multiple sequence alignment, and may be only a fragment of an entire protein sequence.

1. Initially, let each sequence form a separate equivalence class, and a leaf in the evolutionary tree. Create a profile for each row in the alignment. This is accomplished using our standard method: obtain a posterior estimate over the amino acids at each position by combining the observed counts with a Dirichlet mixture prior.
2. While the number of classes in the partition is greater than 1, do:
   (a) Compute the total relative entropy (TRE) (Equation 1) between every pair of profiles.
   (b) Find the pair having the lowest TRE.
   (c) Replace these two classes with a single class that combines the counts at each position. Form an internal node to represent this new class, adding edges to the nodes (or leaves) representing the classes joined. This reduces the number of classes in the partition by 1.
   (d) Estimate the number of independent observations in the new class, and weight the sequences accordingly.[1]
   (e) Create a profile of the expected amino acids at each position for the new class using the weighted counts, in combination with a Dirichlet mixture prior.

Identifying subfamilies in the data

As in classification problems in general, we want to partition the sequences into classes such that the sequences within each class have a high degree of similarity to each other, but not so high as to have a trivial partition with each class containing a single protein. Accordingly, we want to find a partition that minimizes the number of classes while maximizing the similarity among sequences within each class.

At this juncture, two points are relevant: (1) Even when a phylogenetic tree accurately reflects the functional hierarchy in the data, no single cut of the tree into subtrees may be necessarily more correct than another. Each cut may simply reflect a different, and potentially equally correct, decomposition of the sequences into subfamilies. (2) The number of ways to cut a tree with n leaves grows rapidly in n, making an examination of all possible cuts infeasible for protein superfamilies where n can be in the hundreds.

[1]This estimation of the number of independent observations in the data is critical in Bayesian methods, as the formula used to compute the posterior estimate of the expected amino acid distribution at a position will converge on the frequencies in the observed amino acids as the observations increase.

Relative entropy between distributions		
	Sim. AA Types	Diff. AA Types
Conserved Dist.	Low (or zero)	Large
Mixed Dist.	Low (or zero)	Moderate

Table I. Interaction between relative entropy and amino acid distributions in profiles in Bête tree estimation. This table shows the relative entropy between two distributions for four types of cases: conserved distributions preferring similar amino acid types, conserved distributions preferring different amino acid types, mixed distributions preferring similar amino acid types, and mixed distributions preferring different amino acid types. The symmetrized relative entropy (Equation 1, fixed for a single column c, instead of summing over all columns) is largest when two distributions are conserved for different amino acids, especially when the amino acids are of different types. This value is smallest when the distributions are conserved for the same amino acid. In the case where the two distributions are mixed (not showing a strong preference for a particular amino acid) there are two possibilities. If both mixed distributions prefer similar types of amino acids (polar, for example), the relative entropy is small, but if the mixed distributions are of different types (e.g., one prefers polar amino acids, while the other prefers non-polar amino acids), the relative entropy will be larger, but still of moderate size. Because the Dirichlet mixture densities tend to generalize the posterior estimates toward the background distribution in the case where mixed amino acids are observed, the relative entropy between two different mixed distributions is larger than when the mixed distributions prefer similar amino acids, but is still not very large. When a conserved distribution disagrees with a mixed distribution, the relative entropy is larger than when two mixed distributions disagree, but not as large as when two conserved distributions disagree.

To simplify this search for an optimal cut of the tree into subtrees, we only examine a subset of all possible cuts: those obtained during the agglomeration used to construct the tree. Since each iteration of the agglomeration algorithm induces a new set of multiple alignments, one for each class, we can compute the encoding cost of each multiple alignment using Dirichlet mixture densities. In addition, we will measure the model complexity as the cost to encode the subfamily labels for the sequences in the family.

We define an encoding cost (in bits) for every stage of the agglomeration for a multiple sequence alignment of N sequences and S subfamilies, under a Dirichlet density with parameters Θ, to be

$$N \log_2 S - \sum_c \log_2 \text{Prob}\left(\vec{n}_{c,1} \ldots \vec{n}_{c,S} \,\middle|\, \Theta\right) \qquad (3)$$

where $\vec{n}_{c,i}$ is a vector summarizing the observed amino acids in subfamily i at column c in the multiple alignment. $N \log_2 S$ is the maximum cost (in bits) to specify

the subfamily assignments of the sequences in the alignment. The second half of the cost, encoding the multiple alignments induced by the partition, is minimized when the columns in each subfamily alignment have high probability under the Dirichlet mixture prior with parameters Θ. At termination, that partition of the sequences which gives the minimum encoding cost defines the cut of the tree into subtrees, and the corresponding partition of the proteins into subfamilies.

The two costs to encode the data balance each other: the first seeks to minimize the number of subfamilies, while the second seeks to obtain a decomposition into subfamilies (and corresponding alignments) that maximizes the number of columns containing pure, or very similar, amino acid distributions.

Analysis of SH2 Domains

SH2 domains are particularly interesting to biologists because of their involvement in a variety of intracellular signal transduction pathways. First identified as an important functional motif on the basis of a sequence homology between Src and Fps, currently more than 100 SH2 domains in a variety of organisms ranging from sponge to human have been identified. Mutations of these proteins are implicated in certain diseases and disorders, including diabetes, malignant melanoma and asthma. Because of this, there exists a large body of experimental work on these proteins to determine both the substrate specificity of individual members of the family and identify key binding pocket positions (Waksman *et al.* 1993; Songyang *et al.* 1993). A solved crystal structure (1SPSA) is available for this family, as well as careful analysis of both the complexed and uncomplexed conformations (Waksman *et al.* 1993)[2].

SH2 domains contain three binding pockets: glutamate binding, hydrophobic binding, and phosphotyrosine binding (Waksman *et al.* 1993; Songyang *et al.* 1993). Table III shows the amino acids aligned at each of the binding pocket positions in the molecule for each of the subfamilies in the alignment.

The alignment employed as the basis for this phylogenetic analysis[3] shows a moderately high level of primary sequence diversity. Only four of the 103 columns are perfectly conserved, and the average and minimum pairwise residue identities are as low as 39% and 19%, respectively. This level of evolutionary divergence, in combination with the large number of taxa, is known to be challenging to phylogenetic inference methods (Erdos *et al.* 1997). The substantial experimental biological data available for this family makes it attractive for comparing the relative merits of different tree topologies.

Experiment 1: SH2 domain subfamily identification and analysis

In the first experiment, I constructed phylogenetic trees on all 99 taxa in the alignment, using Bayesian Evolutionary Tree Estimation (Bête), Neighbor-joining from the PHYLIP package (Felsenstein 1997), and Star Decomposition from the MOLPHY suite (Hasegawa and Adachi 1997). Whereas much of the fine-branching tree topologies (tree structure for closely related sequences) were consistent across the different methods, there were disagreements on the coarse-branching order relating whole subtrees, particularly between Star Decomposition and Bête. These differences are explored further in the second experiment later in this paper.

The Bête subfamily decomposition produced 15 subfamilies (see Table II), of which three were singletons (Shc_human, Srk3_spola, and Vav_human). All three singletons have features that differentiate them from the other proteins in the data. Shc_human has a five-residue deletion at positions 58-62, at the convergence of the three binding pockets, and has been noted to have significant differences in structure from other SH2 domains (Mikol *et al.* 1995). Srk3_spola is a fragment (deleting the first 60 residues), and comes from freshwater sponge– a very primitive metazoan. Vav_human deletes positions 91-97 centered around the hydrophobic binding pocket, and has a distinctly different amino acid signature at the remaining binding pocket positions.

Only two sequences received novel classifications that were not confirmable by either SwissProt references or by literature search: Src2_drome (placed with the Btk subfamily), and Nck_human (placed with Sem5, Drk, and Grb2). These two cases are discussed below.

Assignment of Src2_drome to Btk subfamily At the time of the original Bête analysis (spring, 1997), Src2_drome was noted in SwissProt as belonging to the Src subfamily; Src2_drome's placement in the Btk subfamily by the method necessitated additional verification.

An examination of the alignment of the SH2 domain of this protein to both the Src subfamily members and the Btk subfamily members showed it to be much more similar to the Btk subfamily (between 44-53% pairwise residue identity) than to the Src subfamily (between 27-36% residue identity). Based on these observations, the subfamily assignment of Src2_drome has been changed in SwissProt from the Src to the Btk subfamily (Amos Bairoch, personal communication).

[2] An introduction to the essentials of structure and function for this diverse family can be found on the World Wide Web, at http://expasy.hcuge.ch/cgi-bin/get-prodoc-entry?PDOC50001.

[3] The alignment used in these experiments was obtained by reestimating HSSP alignment 1sps.hssp (Chain A) for sequence homologs to Src_rsvsr, using HMM methods. The alignment, reordered to reflect the proximity of sequences in the phylogenetic tree, and the tree produced by Bête, are available by anonymous ftp from ftp.cse.ucsc.edu, at pub/protein/phylogeny.

Subfamilies identified in SH2 domains		
Members	Notes	Size
Nck/Drk/Sem5/Grb2	See caption	5
Src	Tyrosine protein kinases (EC 2.7.1.112) Src/Src1/Src2_xenla/Srcn/Fgr/Fyn/Yrk/Yes/Frk/Stk/Blk/Lyn/Hck/Lck/Srk1/Srk2/Srk4 All except Srk noted to be members of Src subfamily in SwissProt	43
Abl/Abl1/Abl2	Tyrosine-protein kinase Dash/Abl (EC 2.7.1.112)	7
Tec/Btk/Itk/Txk Src2_drome	Btk Subfamily Tyrosine-protein kinase (EC 2.7.1.112) in SwissProt (except Src2_drome) Tyrosine-protein kinase Src28c (EC 2.7.1.112) Tec: hematopoietic cell lines including myeloid, B-, and T-Cell lineages. Btk: (B Cell progenitor kinase) Itk: (T-Cell-specific kinase) Txk: (Resting lymphocyte kinase)	9
ZA70/SYK	ZA70/SYK subfamily noted in SwissProt Tyrosine-protein kinase (EC 2.7.1.112)	4
PTNB/PTN6/CSW	PTNB: Protein-tyrosine phosphatase (EC 3.1.3.48)	5
PIP4/PIP5	1-Phosphatidylinositol-4,5-bisphosphate phosphodiesterase Gamma 1 and 2 (EC 3.1.4.11)	5
Crk/Crkl/Gagc_avisc	Crk and Crk-like (CRKL) with avian virus GAGC_AVISC Crk: Proto-oncogene C-CRK (P38) Crkl: Crk-like protein Gagc_avisc: P47(GAG-CRK) Protein. Avian virus	4
CSK/CTK	Tyrosine-protein kinase (EC 2.7.1.112) SwissProt note: belong to Csk subfamily	7
GTPA	Gtpase-activating protein (GAP) (RAS P21 Activation)	2
Fer	Proto-oncogene, non-receptor tyrosine kinase (EC 2.7.1.112)	1
P85A/P85B	Phosphatidylinositol 3-kinase regulatory alpha and beta subunit	4
Vav_human	SwissProt notes: Function: probable exchange factor for a small RAS-like GTP-binding protein. Tissue specificity: widely expressed in hematopoietic cells but not in other cell types.	1
Srk3_spola	Tyrosine-protein kinase (EC 2.7.1.112) (Fragment) - (Freshwater sponge)	1
Shc_human	SHC transforming proteins	1

Table II. SH2 domain subfamilies identified by Bête. See section *Assignment of Nck_human to Drk/Sem5/Grb2 subfamily* for additional information on the Nck/Drk/Sem5/Grb2 subfamily.

Active site positions in SH2 domains																
Bind. Pocket	P	P	P	P	P	P	G	G	GH	GP	P	PH	H	H	H	H
Position	11	31	33	34	35	36	56	57	**58**	59	61	70	71	**86**	**92**	**93**
Nck	R	R	S	E	S	S	K	H	F	K	L	I	G	Y	-	-
Drk/Sem/Grb	R	R	C	E	S	S A	Q	H	F	K	L	L	W	H	-	-
Src	R	R	S	ED	TSH	*	KR	H	Y	KR	RK	IVL	TSA	Y	G	L
Abl	R	R	S	E	ST	S	YF	H	Y	R	NS	VI	TS	H	G	L
Btk	R	R	S	SR	*	-	KR	H	Y	HVQ	KC	ILV	ATS	H	G	L
ZA70/SYK	R	R	R	KD	EN	-	YL	H	Y	LR	SD	I	P	LY	G	L
PTNB/PTN6	G	R	S	LQ	S	QKH	T	H	IV	KM	MR	V	G	FY	-	-
PIP4/PIP5	R	R	S	E	T	F	Q	H	C	R	HR	L	T	Y	E	F
CRK	R	R	S	SG	TS	CSI	S	H	Y	I	N	I	G	Y	W-	D-
CSK/CTK	RG	R	S	TA	NR	HY	EI	H	Y	R	IML	I	D	Y	GA	LI
GTPA	R	R	S	D	R	R	N	H	F	R	I	I	G	Y	-	-
FER	R	R	S	H	G	K	R	H	F	I	Q	R	F	Y	-	-
P85	R	R	S	S	K	-	K	H	C	V	YN	F	A	Y	AS	L
Vav_human	R	R	R	V	K	D	K	H	V	K	M	I	T	Y	-	-
Srk3_spola	-	-	-	-	-	-	-	-	-	-	R	M	A	Y	G	L
Shc_human	R	R	S	T	T	T	K	H	-	-	-	R	T	H	P	I

Table III. Residues in binding pockets of SH2 domains for subfamilies identified using Bayesian Evolutionary Tree Estimation. G=glutamate-binding pocket, P=phosphotyrosine-binding pocket, and H=hydrophobic binding pocket. Positions 58 (βD5, in the standardized notation of Songyang (Songyang *et al.* 1993) and others), 92 and 93 (in **boldface**) are noted in the literature as being the most crucial for determining phosphopeptide specificity (Songyang *et al.* 1993). Note that column 58 is virtually perfectly conserved within each subfamily (the exception being subfamily Ptnb/Ptn6, which has a conservative I-V substitution), illustrating the potential use of this method to flag possible binding pocket positions for experimental verification. Starred (*) columns show several residues aligned by the subfamily in question.

Assignment of Nck_human to Drk/Sem5/Grb2 subfamily

Drk, Grb2 and Sem5 are noted to be functional homologs with identical SH2/SH3 architectures (Stern *et al.* 1993). Nck has not previously been included in this subfamily, apparently due to different orderings of their SH2 and SH3 domains: Nck is composed of three SH3 domains followed by one SH2 domain, whereas Grb2, Drk and Sem5 are all composed of a single SH2 domain sandwiched between two SH3 domains. In the alignment employed for the analysis, the SH2 domain of Nck_human has moderate pairwise residue identities to other SH2 domains – varying from a low of 24% to members of the Btk subfamily, to a high of 43.9% to Drk_drome. Second highest in pairwise residue identity, at 40.5%, to Nck is Srk1_spola, placed in the Src subfamily in this analysis. Sem5_caeel and Grb2_human follow immediately, with 39.8 and 38.6% pairwise identities respectively.

Interestingly, although the pairwise residue identities between Nck and Drk, and between Nck and Srk1_spola are similar overall (43.9% and 40.5% respectively), the difference in pairwise residue identities at the more conserved binding pocket positions is dramatic. Nck and Drk are identical at 12 out of 16 binding-pocket positions (two of which involved deletions), while Nck and Srk1 are identical at only 6 out of 16 of the positions. Significantly, for the three positions noted in the literature as being the most important for determining phosphopeptide specificity (columns 58, 92 and 93 in the alignment (Songyang *et al.* 1993; Waksman *et al.* 1993)) Nck is identical to Drk, Sem5 and Grb2, while Srk1 disagrees at each position (see Table IV).

Two hydrophobic binding-pocket positions (alignment columns 71 and 86) show non-agreement between Nck and Drk, Sem5 and Grb2, which at first glance might make one question the assignment of Nck to this subfamily. However, the bulky tryptophan aligned at position 71 by all except Nck in this subfamily appears to close up the binding pocket, presumably rendering this pocket less important for phosphopeptide specificity among Drk, Sem5 and Grb2 (Waksman *et al.* 1993; Songyang *et al.* 1993).

In addition to similarities at these binding pockets, there are functional similarities among these proteins. Grb2, Nck and Drk are all noted as being adaptor proteins; all bind to growth factor receptors, and are involved in RAS activation (Lu *et al.* 1997).

Examination of the multiple of alignment of Nck with other members of this subfamily revealed two regions where minor hand-editing would improve the pairwise residue identities at non-binding pocket positions (see Figure 1). This increased the pairwise residue identity between Nck and Drk to 48%.

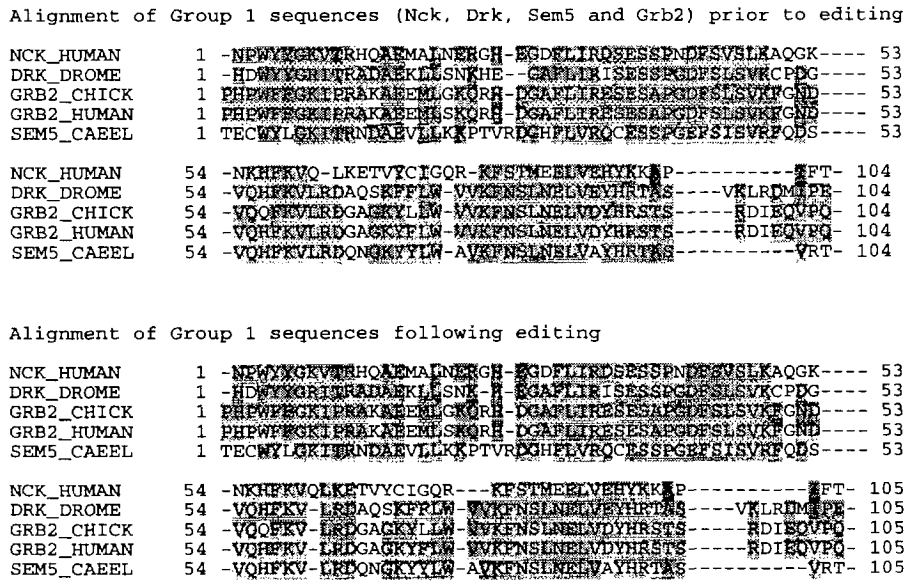

Fig. 1. SH2 Domains: Alignment of Nck and subfamily members, before (above) and after hand-editing regions 23-25 and 62-64 to increase sequence similarity (below). The alignment used to infer the evolutionary tree was the unedited version.

Amino acids in binding pockets in SH2 domains																
Binding Pocket	P	P	P	P	P	P	G	G	GH	GP	P	PH	H	H	H	H
Pos.	11	31	33	34	35	36	56	57	58	59	61	70	71	86	92	93
Group 1																
Drk/Sem5/Grb2	R	R	C,S	E	S	S, A	Q	H	F	K	L	L	W	H	-	-
Nck	R	R	S	E	S	S	K	H	F	K	L	I	G	Y	-	-
Group 2:																
Src subfamily	R	R	S	E	T	T	K	H	Y	K	R	I	T	Y	G	L
Group 3:																
Fgr/Fyn/Yrk/Yes	R	R	S	E	T	T	K	H	Y	K	R	I	T	Y	G	L
Group 4:																
Lyn	R	R	S	E	T	L	K	H	Y	K	R	I	S	Y	G	L
Hck	R	R	S	E	T	T	K	H	Y	K	R	I	S	Y	G	L
Lck	R	R	S	E	S,T	S,T	K	H	Y	K	R	I	S	Y	G	L
Blk	R	R	S	E	S	N	K	H	Y	K	R	I	S	Y	G	L
Group 5:																
Srk	R	R	S	D,E	T	T	R	H	Y	R	R,K	V	T	Y	G	L
Stk	R	R	S	E	T	T	K	H	Y	R	R	I	T	Y	G	L
Frk	R	R	S	E	S	Q	K	H	Y	R	K	L	T	Y	G	L
Group 6:																
Src1_drome	R	R	S	E	H	N	K	H	Y	R	K	I	A	Y	G	L

Table IV. Residues in binding pockets for subgroups found in either the Bête or Star Decomposition trees for selected SH2 domains. G=glutamate-binding pocket, P=phosphotyrosine-binding pocket, and H=hydrophobic binding pocket. Positions 58, 92 and 93 are noted in the literature as being the most crucial for determining phosphopeptide specificity. Dashes indicate deletions in the alignment. Analysis of binding pocket positions for these groups shows a hierarchy of similarity among the groups: Groups 2, 3, and 4 are highly similar in these positions. Groups 2 and 3 are perfectly identical at all positions, while Group 4 agrees at 13 out of 16 positions with Groups 2 and 3. Group 5 is somewhat more variable (in particular, at position 59, where Group 5 sequences substitute R for the otherwise conserved K). These groups also cluster separately with respect to tissue expression. Group 3 proteins (Fgr, Fyn, Yrk and Yes) are expressed primarily in neural tissues, whereas Group 4 proteins (Lyn, Hck, Lck, and Blk) are expressed primarily in B and T lymphoid cells. Groups 1 and 6 have distinguishing features which set them apart from the other groups. Group 1 aligns F intead of consensus Y at position 58, which has been identified experimentally as being the most important position for determining phosphopeptide specificity, and also aligns L at 61, instead of the consensus positively charged residue all other groups align. Sequences in this group also delete positions 92 and 93, the other two crucial positions for determining substrate specificity, which shows a conserved GL motif for all other groups. Group 6 consists only of Src1_drome. This protein aligns H at 35, where all other groups align T or S, R at 59 (disagreeing with all except group 5), and A at 71, which has T or S in all groups save Group 1.

Experiment 2: Tree topology comparisons on sequences in Src and Nck/Drk/Sem5/Grb2 subfamilies

For a more in-depth analysis, I selected two subfamilies identified in the first experiment which were found in adjacent subtrees in the tree estimated by Bête: 47 sequences from the Src and Nck/Drk/Sem5/Grb2 subfamilies. The multiple alignment of these proteins was extracted from the larger alignment, and used as the basis for estimating trees using Star Decomposition and Maximum Likelihood from the MOLPHY software suite (Hasegawa and Adachi 1997), Neighbor Joining and Parsimony from the PHYLIP software suite (Felsenstein 1997), and Bête. Because Maximum Likelihood examines all tree topologies (a quantity that grows exponentially in the number of sequences), a subset of only 11 sequences were used in the ML analysis. For comparison purposes, the same subset was used in the Parsimony analysis. The tree produced by Neighbor Joining was quite similar to that produced by Bête; for space considerations, only the Bête tree is shown of the two.

From the two trees inferred using Star Decomposition and Bête from the alignment of 47 proteins, several subgroups were identified, and similarities at binding pocket positions analyzed (see Table IV). Groups 1, 2 and 4 were separated into distinct subtrees by both methods; proteins in groups 3 and 5 are gathered into a subtree by one but not both methods. Group 6, consisting of Src1_drome, is a singleton in both trees.

The similarities between groups at binding pocket positions as shown in Table IV, are reflected in the tree topologies produced by Bête (Figure 3), and Neighbor-Joining (data not shown) but not by Star Decomposition (Figure 2).

For example, Star Decomposition places groups 2 (Src) and 3 (Fgr/Fyn/Yrk/Yes), which are indistinguishable in the binding sites, at opposite ends of the tree, while Bête

links them most closely, placing them in a subtree. Star Decomposition also links groups 1 (Nck et al) and 2 (the Src subfamily) most closely, while these two groups are maximally separated in the Bête tree. Although the Star Decomposition tree is unrooted, allowing us to pick the root to resolve ambiguities, it is not possible in this case to pick the root so that groups 2 and 3 are in the same subtree.

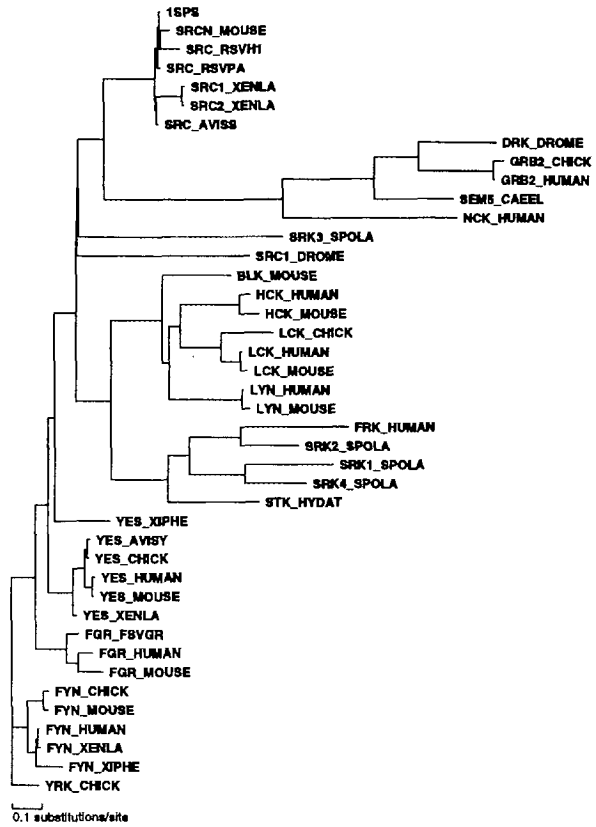

Fig. 2. Star Decomposition tree for SH2 domains used in Experiment 2. This tree contains taxa from six groups whose binding-pocket positions and functional roles are discussed in Table IV. Groups 2 (Src, at top) and 3 (Fgr/Fyn/Yrk/Yes, at bottom) are identical in the binding pocket positions, and ought to be placed adjacent in a tree. 1SPS is the PDB identifier for the SwissProt sequence Src_rsvsr. Star Decomposition software was obtained from the MOLPHY software suite (Hasegawa and Adachi 1997)

Conclusions

A novel method for estimating phylogenetic trees on protein superfamilies, incorporating Bayesian and information-theoretic methods, has been presented. This method identifies, from the primary sequence alone, residues that are conserved within subfamilies for functional or structural reasons, and drives the tree topology

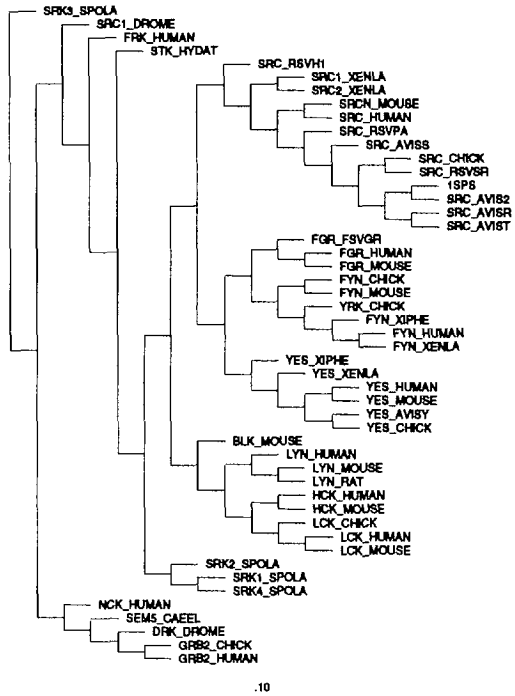

Fig. 3. Bête tree for SH2 domains used in Experiment 2. The distances between the taxa in this tree are proportional to the differences in their residues aligned in binding pocket positions, as shown in Table IV. Groups 2 (Src) and 3 (Fgr/Fyn/Yrk/Yes) are identical in the binding pocket positions, and placed adjacent in this tree. The next closest subtree to these two contains Group 4 proteins (Lyn/Hck/Lck/Blk), which are identical at 13 out of 16 binding pocket positions. The last group to be joined (before the fragment Srk3_spola) contains Group 1 sequences (Nck/Drk/Sem5/Grb2), which show significant differences in binding pocket positions from others in the alignment used as input. 1SPS is the PDB identifier for the SwissProt sequence Src_rsvsr. Edge lengths drawn are all unit length, and do not correspond to the distance measurement computed to infer the tree.

estimation to conserve these residues within subtrees. The information-theoretic method for obtaining a cut of the tree into subtrees produces a classification of the sequences into subfamilies that reflect the functional similarity among the proteins.

Applied to SH2 domains in this paper, this method is shown to produce a tree topology that clusters together into subtrees proteins which are similar at the binding-pocket positions. The cut of the tree into subtrees, and thus subfamilies, reveals subfamily-specific conservation at these positions, suggesting the applicability of this method as a predictive tool for further experimental val-

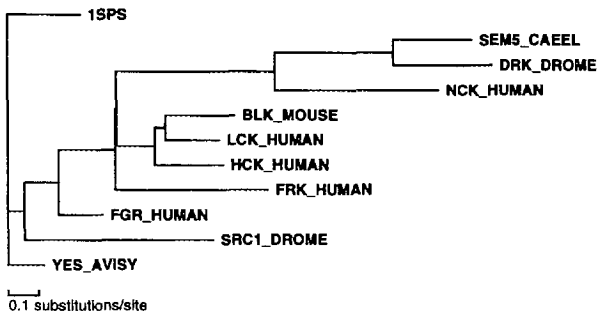

Fig. 4. Maximum Likelihood tree for subset of SH2 domains used in Experiment 2. This tree contains taxa from six groups whose binding-pocket positions and functional roles are discussed in Table IV. Fgr and Yes are in Group 3, which are identical in binding pocket residues, and so ought to be adjacent in the tree. In this tree, however, Src1_drome (from Group 6) is interspersed between Fgr and Yes, making it impossible to choose a root to obtain a monophyletic Group 3. It is hard to know why the ML tree chose this topology; the pairwise residue identity between Fgr_human and Yes_avisy is 75%, and 65% between Fgr_human and 1SPS, a significant increase over the identity between Src1_drome to either Fgr_human (50%) or Yes_avisy (53%). For computational reasons, only 11 sequences were used to estimate this tree topology. The program used came from the MOLPHY suite (Hasegawa and Adachi 1997).

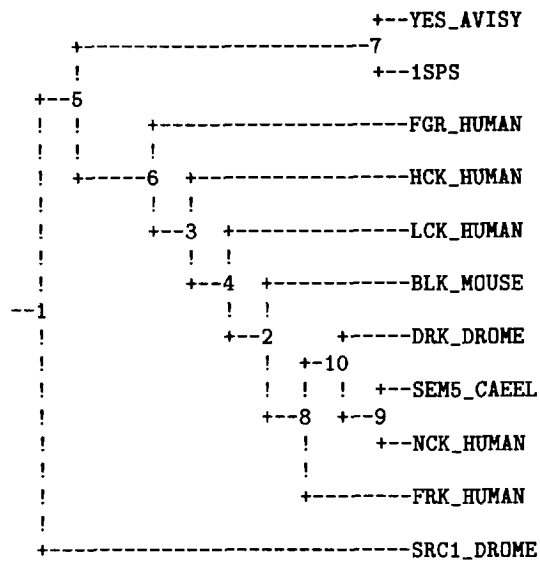

Fig. 5. Parsimony tree estimated using protpars from the PHYLIP software suite (Felsenstein 1997). A reduced set of sequences was chosen for comparison with the Maximum Likelihood tree in Figure 4, and for computational efficiency. As in the ML tree, it is not possible to choose a root to obtain a monophyletic group 3 (Fgr/Fyn/Yrk/Yes).

idation. This method resulted in a change in the SwissProt annotation for Src2_drome, and suggests a new subfamily classification for Nck_human.

Acknowledgements

This work was funded, in part, by a graduate research fellowship from the National Science Foundation, and by a fellowship from the Program in Mathematics and Molecular Biology. David Haussler made important contributions to the development of the method. I also want to thank the anonymous referees for their careful reading of the text and valuable recommendations, and Tandy Warnow for editorial advice and scientific discussion.

References

Bailey, Timothy L. and Elkan, Charles 1995. The value of prior knowledge in discovering motifs with MEME. In *ISMB-95*, Menlo Park, CA. AAAI/MIT Press. 21–29.

Brown, M. P.; Hughey, R.; Krogh, A.; Mian, I. S.; Sjölander, K.; and Haussler, D. 1993. Using Dirichlet mixture priors to derive hidden Markov models for protein families. In Hunter, L.; Searls, D.; and Shavlik, J., editors 1993, *ISMB-93*, Menlo Park, CA. AAAI/MIT Press. 47–55.

Casari, G.; Sander, C.; and Valencia, A. 1995. A method to predict functional residues in proteins. *Structural Biology* 2:171–178.

Cover, Thomas M. and Thomas, Joy A. 1991. *Elements of Information Theory.* John Wiley and Sons, first edition.

Erdos, P.L.; Steel, M.; Szekely, L.; and Warnow, T. 1997. A few logs suffice to build (almost) all trees. Technical Report 97-71, DIMACS.

Felsenstein, Joe 1996. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods in Enzymology* 266:418–427.

Felsenstein, Joe 1997. Phylip software. http://evolution.genetics.washington.edu/phylip.html.

Hasegawa, M. and Adachi, J. 1997. Molphy 2.2 software. http://dogwood.botany.uga.edu/malmberg/software.html.

Hasegawa, M.; Kishino, H.; and Saitou, N. 1991. On the maximum likelihood method in molecular phylogenetics. *J. Mol. Evol.* 32:443–445.

Karplus, Kevin; Sjölander, Kimmen; Barrett, Christian; Cline, Melissa; Haussler, David; Hughey, Richard; Holm, Liisa; and Sander, Chris 1997. Predicting protein structure

using hidden Markov models. *Proteins: Structure, Function, and Genetics.*

Kuhner, M.J. and Felsenstein, Joe 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol Biol Evol* 11:449–468.

Lichtarge, O; Bourne, H.R.; and Cohen, F.E. 1996. An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*. 257:342–358.

Lu, W; Katz, S; Gupta, R; and Mayer, BJ 1997. Activation of Pak by membrane localization mediated by an SH3 domain from the adaptor protein Nck. *Curr Biol.*

Miklos, G.L. and Rubin, G.M. 1996. The role of the genome project in determining gene function: insights from model organisms. *Cell* 86:521–529.

Mikol, V; Baumann, G; Zurini, MG; and Hommel, U 1995. Crystal structure of the SH2 domain from the adaptor protein shc: a model for peptide binding based on x-ray and nmr data. *J Mol Biol* 254:86–95.

Sjölander, K.; Karplus, K.; Brown, M. P.; Hughey, R.; Krogh, A.; Mian, I. S.; and Haussler, D. 1996. Dirichlet mixtures: A method for improving detection of weak but significant protein sequence homology. *CABIOS* 12(4):327–345.

Sjölander, Kimmen 1997. *A Bayesian-Information Theoretic Method for Evolutionary Inference in Proteins.* Ph.D. Dissertation, University of California Santa Cruz.

Sjölander, K. 1998. Bayesian evolutionary tree estimation. *Mathematical Modelling and Scientific Computing.* To appear.

Songyang, Z; Shoelson, SE; Chaudhuri, M; Gish, G; Pawson, T; Haser, WG; King, F; Roberts, T; Ratnofsky, S; Lechleider, RJ; and al, et 1993. SH2 domains recognize specific phosphopeptide sequences. *Cell* 72:767–778.

Stern, MJ; Marengere, LE; Daly, RJ; Lowenstein, EJ; Kokel, M; Batzer, A; Olivier, P; Pawson, T; and Schlessinger, J 1993. The human Grb2 and drosophila Drk genes can functionally replace the caenorhabditis elegans cell signaling gene Sem-5. *Mol Biol Cell* 4:1175–1188.

Tateno, Y.; Takezaki, N.; and Nei, M. 1994. Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. *Mol Biol Evol* 11:449–468.

Tatusov, Roman L.; Altschul, Stephen F.; and Koonin, Eugen V. 1994. Detection of conserved segments in proteins: Iterative scanning of sequence databases with alignment blocks. *PNAS* 91:12091–12095.

Waksman, G; Shoelson, SE; Pant, N; Cowburn, D; and Kuriyan, J 1993. Binding of a high affinity phosphotyrosyl peptide to the Src SH2 domain: crystal structures of the complexed and peptide-free forms. *Cell* 72:779–790.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* 39:306–314.