Statistical Limits to the Identification of Ion Channel Domains by Sequence Similarity

Anthony A. Fodor and Richard W. Aldrich

Department of Molecular and Cellular Physiology, Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, CA 94305

The study of ion channel function is constrained by the availability of structures for only a small number of channels. A commonly used bioinformatics technique is to assert, based on sequence similarity, that a domain within a channel of interest has the same structure as a reference domain for which the structure is known. This technique, while useful, is often employed when there is only a slight similarity between the channel of interest and the domain of known structure. In this study, we exploit recent advances in structural genomics to calculate the sequence-based probability of the presence of putative domains in a number of ion channels. We find strong support for the presence of many domains that have been proposed in the literature. For example, eukaryotic and prokaryotic CLC proteins almost certainly share a common structure. A number of proposed domains, however, are not as well supported. In particular, for the COOH terminus of the BK channel we find a number of literature proposed domains for which the assertion of common structure based on common sequence has a nontrivial probability of error.

The Journal of General Physiology

INTRODUCTION

If the sequence similarity between a protein of interest and a protein of known structure is strong, we have a high degree of confidence that the two proteins share a common fold. In this case, experiments can focus on detailed molecular mechanisms, at the level of individual residues, that may be similar or different between the two proteins. If the sequence similarity is weak, we are less certain of common structure, and a great deal of experimental evidence is required to broadly establish similarity of mechanism. If we are to judge the appropriate level of experimental inquiry for a given protein, therefore, we must be able to gauge our confidence in the existence of a common fold based on the strength of sequence similarity.

In the literature, the degree of sequence similarity between two proteins is often measured by percent identity. If the percent identity is above ~35%, it can reliably be asserted that two proteins share a common fold (Rost, 1999). It is, however, an unfortunate feature of proteins that many have percent identities below 30% yet still share a common fold. Indeed, many proteins that have common folds, and many that do not, have percent identities between 20 and 30%(Brenner et al., 1998; Rost, 1999). This ambiguous 20–30% sequence identity range is often referred to as the "twilight zone."

This problem of resolving proteins in the "twilight zone" is particularly acute in the study of ion channels. Even despite considerable recent progress (Long et al., 2005a,b), electrophysiological experiments are usually performed on eukaryotic channels, while the majority of solved channel structures are from prokaryotes. The great evolutionary distance between prokaryotes and eukaryotes means that comparisons between bacterial and vertebrate channels rarely hits the 30% identity mark. Nonetheless, it is widely assumed in the literature that eukaryotic and prokaryotic channels share common folds. In this paper, we exploit recent progress in structural genomics to quantify the reliability of these sorts of assumptions. We find that for a diversity of eukaryotic channels, folds can indeed be assigned with very low false positive rates. We also explore cases in the channel literature where there appears to be little sequence-based evidence for domains that have been proposed based on sequence similarity.

MATERIALS AND METHODS

The HMMer model library (Gough et al., 2001) (version 1.67) was downloaded from the Superfamily web site (http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/). For query sequences, we used the Astral database (Brenner et al., 2000; Chandonia et al., 2002; Chandonia et al., 2004), which provides easily parsable files containing the sequences, and SCOP (structural classification of proteins) superfamily assignments, of domains from the PDB. We used the Astral database filtered at 40% sequence identity for

Correspondence to Anthony Fodor: anthony.fodor@gmail.com

A. Fodor's present address is Department of Bioinformatics/Computer Science, University of North Carolina, Charlotte, 9201 University City Blvd., Charlotte, NC 28223.

R. Aldrich's present address is Section of Neurobiology, University of Texas at Austin, Austin, TX 78712.

The online version of this article contains supplemental material.

Abbreviations used in this paper: BK, large conductance calcium-activated potassium; HMM, hidden Markov model; RCK, regulator of potassium conductance; SCOP, structural classification of proteins.

our query sequences. For each sequence in the query set, we used the hmmsearch program of the HMMer package (http://hmmer.wustl.edu/, version 2.3.2 with the default parameters under Cygwin and OS X) to run the query sequence over each of the 9,939 profiles in the Superfamily database (Fig. 1). We refer to the sequence that was used to seed each Superfamily profile as the "seed sequence."

In compiling our results for estimating false positive rates, we only counted hits where both the query sequence and the seed sequence were present in the ASTRAL database filtered at 40% identity. That is, no protein in either the query set or seed sequence set had >40% sequence identity to any other protein within the query or seed sequence set. In this way, we guarantee that, for our calculation of false positive rates in Fig. 2, large numbers of similar sequences do not distort our results on either the query or the target side. When attempting to find ion channel domains, however, we wished to perform as inclusive a search as possible. Therefore, in creating Figs. 3–8, we allowed hits against any Superfamily profile.

We scored a hit as "correct" when the query and the seed sequence belonged to the same SCOP superfamily. We scored a hit as "incorrect" when the query sequence and the seed sequence belonged to different SCOP superfamilies and different SCOP folds. We ignored hits in our false positive estimates in which the query sequence and seed sequence belonged to different SCOP superfamilies but the same SCOP fold. By these criteria, we generated 107,425 hits of which 67,158 were marked correct and 40,267 were incorrect.

We note that our metric of false positive rate is more discriminating than the "e-value" scores that the HMMer program generates. An e-value is a commonly used metric that is defined as the number of hits that one would expect if the search were performed using random sequences. Even when our false positive rates were ~80%, we still see e-values of <0.0001 (see online supplemental material, Tables S1–S6, available at http://www.jgp.org/cgi/content/full/jgp.200509419/DC1). This discrepancy may represent HMMer over-stating the significance of its searches or be a feature of the construction of the Superfamily models. Alternatively, it may reflect that proteins with significant sequence similarity due to common ancestry may no longer share a common fold. Resolving this question is beyond the scope of this paper.

Code wrapped around the HMMer distribution was used to perform these analyses and generate all the figures in this paper. Java code and instructions for reproducing Figs. 2–8 in the paper are available upon request.

Online Supplemental Material

The online supplemental material (available at http://www.jgp.org/cgi/content/full/jgp.200509419/DC1) shows the results of running ion channel query sequences against the 9,939 Superfamily profile HMMs representing protein domains of known structure (Tables S1–S6).

RESULTS

Estimation of False Positive Rates

We have a large set of ion channels for which we do not have direct structures. We wish to know how much confidence we can place in the assignment of structure to channels in this set based on sequence similarity to protein domains of known structure. That is, we wish to ask what is the degree that two sequences must be related before we can become confident that they share the same fold. To understand the relationship between sequence similarity and common fold, we can turn to the large number of proteins whose crystal structures have been solved. For each of the domains in the PDB, we can ask: if we did not know the structure, and had to guess the structure based only on sequence similarity to another protein domain of known structure, how confident could we be of our guess as a function of how closely the sequences of the two proteins are related?

To compute the answer to this question, we turn to three preestablished databases (see Fig. 1). Our first requirement is classification of each structure in the PDB. The SCOP database, which was created with a combination of manual and automated curation, describes each domain in the PDB with a controlled vocabulary (Murzin et al., 1995; Hubbard et al., 1997; Hubbard et al., 1998, 1999; Lo Conte et al., 2000, 2002; Andreeva et al., 2004). The SCOP database defines proteins with a common "fold" as having the same pattern of major secondary structures. This would appear to be the classification level that we are interested in. However, classification efforts in the literature are often at the "superfamily" level (Gough et al., 2001; Madera et al., 2004). The SCOP database defines a superfamily as a set of proteins that, based on structure, have a probable common evolutionary origin. Because the great majority of folds in SCOP (768 of 887) contain only one superfamily, it makes little practical difference whether we work at the superfamily or fold level. Since there is a fairly developed literature on the problem of predicting SCOP superfamily from domain sequence (Gough et al., 2001; Madera et al., 2004; Wistrand and Sonnhammer, 2005), we choose for the rest of this paper to work at the superfamily level.

While the SCOP database gives us an overview of which protein domains belong to the same superfamily, we also require a way to measure the degree to which a query sequence shares sequence similarity with a target sequence of known structure. A very intuitive metric of the similarity between two proteins is percent identity, as can be generated by a global alignment program such as CLUSTAL. It has become apparent, however, that the percent identity metric is a particularly poor way of assessing whether two proteins share the same fold (Rost, 1999). Scores produced by local alignment programs such as BLASTP perform with much greater power and sensitivity than does percent identity when used as a metric to determine if two distantly related proteins belong to the same superfamily (Brenner et al., 1998). Indeed, on the order of half of all proteins that are in the "twilight zone" of 20–30% sequence identity can be clearly resolved by using scores generated by BLASTP instead of percent identity to measure how closely two proteins are related (Brenner et al., 1998). Another weakness of the percent identity metric is that, by its nature, it involves a pairwise comparison between two proteins. A much more powerful approach involves

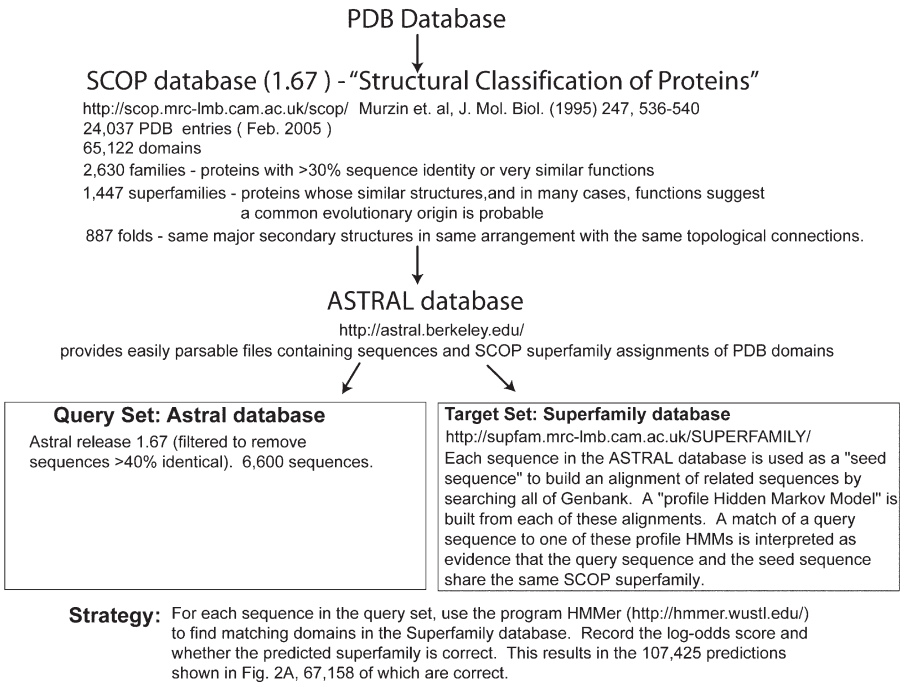

Figure 1. The databases used in this study.

building an alignment of related proteins and comparing the query sequence to this alignment (Rychlewski et al., 2000; Gough et al., 2001). In particular, the Superfamily database (Gough et al., 2001; Madera and Gough, 2002; Madera et al., 2004) has proven to be a powerful tool for determining if two proteins share the same fold.

The Superfamily database is a collection of searchable profiles that represent all proteins of known structure (Gough et al., 2001). The construction of the Superfamily database starts with sequences from the PDB filtered at 95% sequence identity to remove close duplicates (Fig. 1; for details of the Superfamily technique see Gough et al., 2001). A sequence from each PDB domain, representing a known SCOP superfamily, is used as a "seed sequence" to build a multiple sequence alignment by searching all the proteins in Genbank/EMBL/DDBJ for related sequences. Next, a hidden Markov model (HMM) is generated that describes the probability of finding each residue at each position in the alignment. This model, called a "profile HMM," is then used to again search all the proteins in Genbank/EMBL/DDBJ for matching proteins. This leads to a new alignment with an increased number of sequences. The new alignment is used to generate a new profile HMM, which in turn is used to build a larger alignment. This recursive search is repeated a set number of times (see Gough et al., 2001, for details). The final profile HMM derived from this process can be used to estimate how well a given query sequence matches the SCOP superfamily of the seed sequence. Given a sequence S, and a Superfamily profile model M, we can use the HMMer software package (http://hmmer.wustl.edu/) to produce a log-odds score indicating how well M fits S. The log-odds score is defined (see HHMer user's guide distributed as part of the HMMer package) as

$$\log_2 \frac{p(S|M)}{p(S|R)},$$

where R is a model describing random sequence. We read this as "the probability of the sequence, given the Superfamily model" divided by "the probability of the sequence, given a model based on random sequence" (see Durbin et al., 1998, for an excellent tutorial on these sorts of statistics). The higher the log-odds score, the higher the probability that S and the seed sequence used to generate M have the same fold.

Our study begins with a demonstration of the relationship between log-odds score and the probability of making an erroneous assignment, defined as assigning a query sequence to an incorrect SCOP superfamily. We start with the ASTRAL database, which provides easily parsable text files that contain sequence and SCOP assignments for every protein domain in the PDB (Fig. 1; see Materials and Methods). We then use the HMMer package to run each sequence in our ASTRAL query set over the 9,939 Superfamily domain models, which represent all known protein structures. For each query sequence matched to each model, we note the log-odds score of the hits (if any) and whether the superfamily of the query sequence is, in fact, the same as the superfamily of the seed sequence. (We ignore cases where the query sequence is identical to the seed sequence.) If the superfamilies are, in fact, different, we mark the hit between the query sequence and the profile model as being in error.

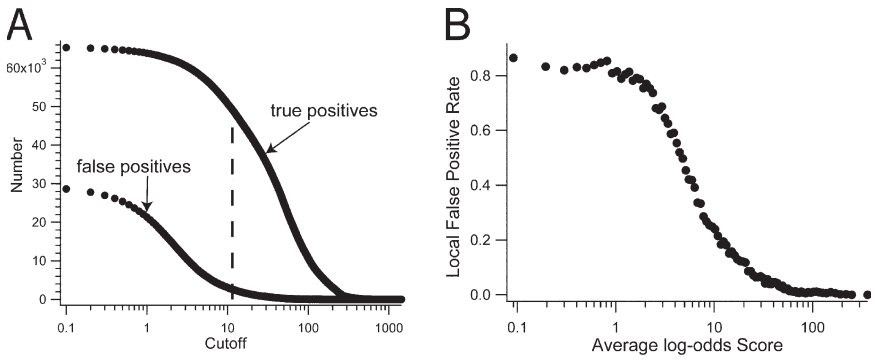

Figure 2. False positive rates of protein domains in the PDB. These data show the results of how well we could ascertain the SCOP superfamily of each domain in our query set based only on sequence similarity (see Materials and Methods). (A) The number of false positives and true positives in our dataset that had log-odds scores greater than or equal to the value shown on the x-axis. (B) The same data used to generate A rearranged to show the false positive rate as a function of log-odds score. Each point is a bin in a histogram representing 1,000 hits. The x-axis shows the average log-odds score of each of the hits in each bin. The y-axis shows the number of incorrect assignments to SCOP superfamily in each bin divided by the total number of hits in each bin.

Fig. 2 A shows the results of the 107,425 hits generated by running the ASTRAL query sequence set against the Superfamily database. Each point in this graph shows the number of true positives or false positives captured when we only include hits that are greater than or equal to the log-odds score given on the x-axis. As we move from right to left on the x-axis, the cutoff score is lowered and we include more hits in our analysis. This increases the number of both false positives and true positives captured. If we take, for example, a cutoff of ∼11.5 (dashed line), we would capture 49,192 true positives and 2,604 false positives, representing an error rate of ∼5%. There are, however, 67,158 total true positives in our dataset. We have, therefore, at a 5% error rate, only captured ∼73% (49,192/67,158) of the true positives in our dataset. This reflects a well known, but unfortunate, difficulty of working with protein sequences: sometimes proteins with little in common, and hence low log-odds scores, nonetheless have the same fold. We therefore need to set a low score cutoff to capture all the proteins with the same fold. A low cutoff, however, also includes many spurious hits of proteins that are not in the same superfamily, which is reflected in Fig. 2 A by the rapid increase in the number of false positives generated when cutoffs below ∼11 are used.

The data used to generate Fig. 2 A can be rearranged to show the probability of making an erroneous assignment as a function of log-odds score. Fig. 2 B is a histogram with each point a bin representing 1,000 hits. The x-axis shows the average log-odds score for the 1,000 hits in each bin. The y-axis shows the number of false positives in each bin divided by the total number of hits in each bin. We see that, below a log-odds score of <∼10 there is a rapid degradation in the quality of the generated predictions. At a log-odds score of <∼1, the prediction that the query sequence is in the same superfamily as the seed sequence is wrong ∼80% of the time.

The data in Fig. 2 B represent the error rates generated by comparing thousands of protein domains of known structure. We can, therefore, use Fig. 2 B as a guide to ask the following: in general, how much confidence do we have that two proteins share the same fold given some level of sequence similarity?

Predicted Results for Kv1.2

To get a sense of how well the methods described above are able to assign probabilities to domain assignments in ion channels, we can compare the predictions of our bioinformatics techniques with the structures of channels that have been directly determined by experiment. We start with the rat Kv1.2 potassium channel, which recently became the first eukaryotic ion channel to have a structure solved for the S1–S6 core region (Long et al., 2005a,b). The version of the Superfamily database we used (1.67) was constructed, however, before this recently solved Kv1.2 channel structure was put into the PDB. This allows us to see how well our methods work in a case where we know, but our database does not, what the correct answer will be. Fig. 3 shows the results of running the Kv1.2 sequence against the 9,939 profiles in the Superfamily database. Each line in Fig. 3 is a hit in which part of the Kv1.2 query sequence matched to a Superfamily profile. The log-odds score of each hit (given in Table S1, available at http://www.jgp.org/cgi/content/full/jgp.200509419/DC1) has been translated to false positive rate using the data in Fig. 2 B. As we move from top to bottom in Fig. 3, the hits against each profile become weaker, and an assertion that Kv1.2 shares a common structure with the seed sequence used to generate that profile is more likely to be in error.

The green lines in Fig. 3 represent hits against profiles with seed sequences that have been assigned to the "voltage-gated potassium channels" SCOP superfamily. We see that the highest scoring, most probable hit for Kv1.2 belongs to this superfamily. The annotations next to the top hits indicate the seed sequence that was used to build each profile (see also Table S1). The prokaryotic KvAP structure was the first channel crystal structure solved that captured a six transmembrane ion channel (Jiang et al., 2003). We see in Fig. 3 that the profile that used KvAP as its seed sequence covers the six

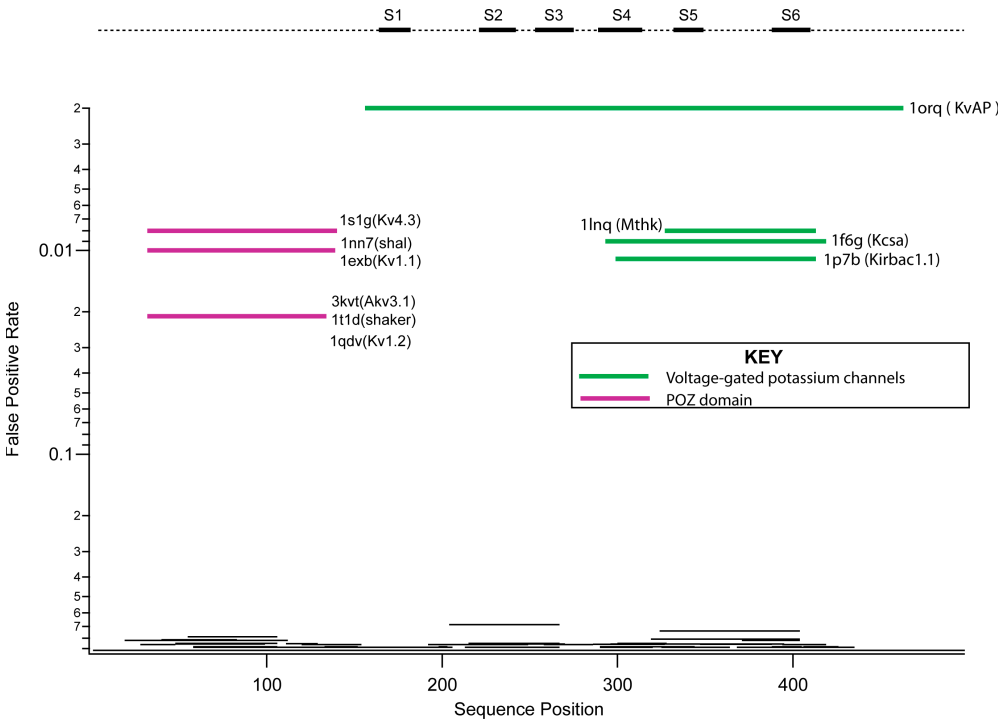

Figure 3. Predicted domains for the rat Kv1.2 channel. The results of running the rat Kv1.2 channel (Genbank/EMBL/DDBJ 52000923) over the 9,939 Superfamily profile HMMs representing protein domains of known structure. The y-axis, the false positive rate, was generated by taking the log-odds score for each hit and converting it to the false positive rate using the data from Fig. 2 B. Annotations next to some of the hits indicate the identity of the seed sequence. The tabular data used to construct this graph are available online as Table S1. S1–S6 regions map predicted transmembrane domains and are taken from McKinnon (1989).

transmembrane domains for Kv1.2 (marked as S1–S6). Based only on these sequence data, we assert with high confidence (a false positive rate <0.003) that the Kv1.2 channel shares a common structure with the crystal structures of the prokaryotic potassium channels. We now know that this is the case for the S5–S6 pore region as the Kv1.2 S5–S6 region can be superimposed with prokaryotic structures (see Fig. 3 in Long et al., 2005a). The situation with the S1–S4 region is more complicated. The prokaryotic KvAP S1–S4 region, when expressed in isolation, can be superimposed on the Kv1.2 S1–S4 region (see Fig. 2 B in Long et al., 2005b). This finding shows that our prediction of common structure for S1–S4 between KvAP and Kv1.2 is not in error. However, in the full-length channel structure of KvAP, regions of the voltage-sensing domains were in a "non-native conformation pulled towards the cytoplasmic side of the pore" (Long et al., 2005b) and are not superimposable with Kv1.2 (see Fig. 2 B in Long et al., 2005b). Resolution of the meaning of the conformation of the S3–S4 region in the full-length KvAP structure remains an area of active enquiry.

The purple lines in Fig. 3 represent hits against profiles whose seed sequence belongs to the "POZ domain" superfamily. This superfamily includes members whose SCOP family is the "tetramerization domain of potassium channels." This structure, known as the T1 domain, has been solved in a number of ion channels, including the Shaker potassium channel (Kreusch et al., 1998) and, as an isolated domain, the Kv1.2 channel itself (Minor et al., 2000). We see that a number of these T1 structures map with high confidence to the NH_2-terminal region. It may seem at first surprising that the Kv1.2 channel, when used as a query sequence, maps to a profile that used Kv1.2 as a seed sequence with a false positive rate >0. This can happen because the seed sequence is used to build a profile that consists of many different related sequences. The log-odds score that is generated is to the entire profile and not just to the seed sequence. Query sequences will therefore not necessarily map with extremely high scores to profiles in which the query sequence was also used as the seed sequence.

The assertions of the presence of the SCOP superfamilies "POZ domain" and "voltage-gated potassium channels" are the only two predictions that we would make for Kv1.2 with a false positive rate of <0.05 (Fig. 3). These results give us some confidence in the ability of these techniques to discriminate true and spurious hits.

Predicted Results for the HCN Channel

Another example of a eukaryotic ion channel for which we have direct structural evidence is the HCN2 channel, which is gated by both voltage and cyclic nucleotides. In the case of this channel, there is a recent crystal structure of the isolated cyclic nucleotide binding domain (Zagotta et al., 2003). The cyclic nucleotide-binding domain in the HCN channel occurs after the S6 region at the end of the channel. When cAMP binds to this region of the channel, the channel is more likely to open (DiFrancesco and Tortora, 1991). Fig. 4 shows the results of running the HCN channel sequence against the Superfamily database. The blue lines in Fig. 4 represent hits against the SCOP superfamily "cAMP-binding

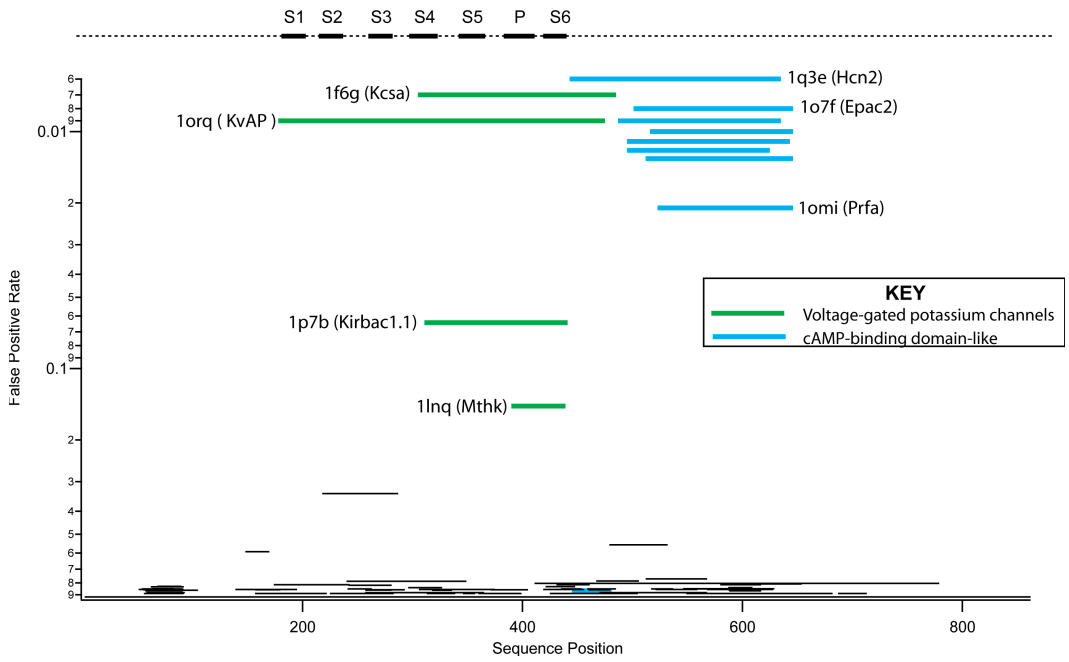

Figure 4. Predicted domains for the mouse HCN2 channel. The results of running the HCN2 channel (Genbank/EMBL/DDBJ 6680189) over the 9,939 Superfamily profile HMMs representing protein domains of known structure. The tabular data used to construct this graph are available online as Table S2. S1–S6 regions map predicted transmembrane domains and are taken from Santoro et al. (1998).

domain-like." As we see, we would have predicted the presence of a cyclic-nucleotide binding domain in this channel based on sequence similarity with a fair number of other previously solved cyclic-nucleotide binding domains (see also Table S2). Because the cyclic-nucleotide binding domain of this channel has, in fact, been solved (Zagotta et al., 2003) and been shown to share a common fold with other cyclic-nucleotide binding domains, we know this prediction is correct. Since the KvAP structure also maps to the HCN channel with a low false positive rate, we can assert with confidence that we know the structure of several key domains of this channel.

Eukaryotic and Prokaryotic CLC Channels Almost Certainly Share a Common Fold

Most functional data from CLC chloride channels has been collected from eukaryotic channels, while structures are available only for prokaryotic homologues (Dutzler et al., 2002; Dutzler et al., 2003). We can use our techniques to ask how appropriate prokaryotic CLC structures are as models for the structure of eukaryotic channels. Fig. 5 shows the results of running CLC-0, from the Torpedo electric ray, against the Superfamily database. As we see, the Torpedo CLC-0 channel maps with a false positive rate of zero to the superfamily profiles that were seeded with the prokaryotic CLC channels. A false positive rate of zero means that there are no structures in the PDB that are as closely related in sequence as the CLCs that do not share the same fold. We therefore have very high confidence in the assertion that prokaryotic and eukaryotic CLC channels have the same fold.

This high confidence in similarity of fold is particularly striking given the recent demonstration that eukaryotic and prokaryotic CLC channels can in fact have different functions. The prokaryotic CLC protein has been demonstrated to be a proton/chloride antiporter (Accardi and Miller, 2004), while the eukaryotic CLC-0 protein is known to be a true chloride channel. This demonstrates that determination of function can depend as much on small-scale molecular interactions as on large domain architectures.

We also see evidence for two CBS domains at the COOH-terminal end of the CLC-0 sequence. The CBS (cystathionine beta synthase) domain has been shown to bind to the adenosyl portion of molecules such as ATP (Scott et al., 2004). Even though none of the individual hits against profiles seeded with CBS domains map with a false positive rate <0.05, there are many CBS domain hits that map to two distinct regions in the COOH terminus with a moderate false positive rate (Fig. 5, green lines). This gives us good confidence that both CBS domains are present in the channel.

Predicting EF-hands in Sodium and Calcium Channels

The potassium and cyclic-nucleotide gated channel genes that we have so far examined create tetrameric channels with four subunits, each one of which consists of a copy of the channel gene. By contrast, sodium and

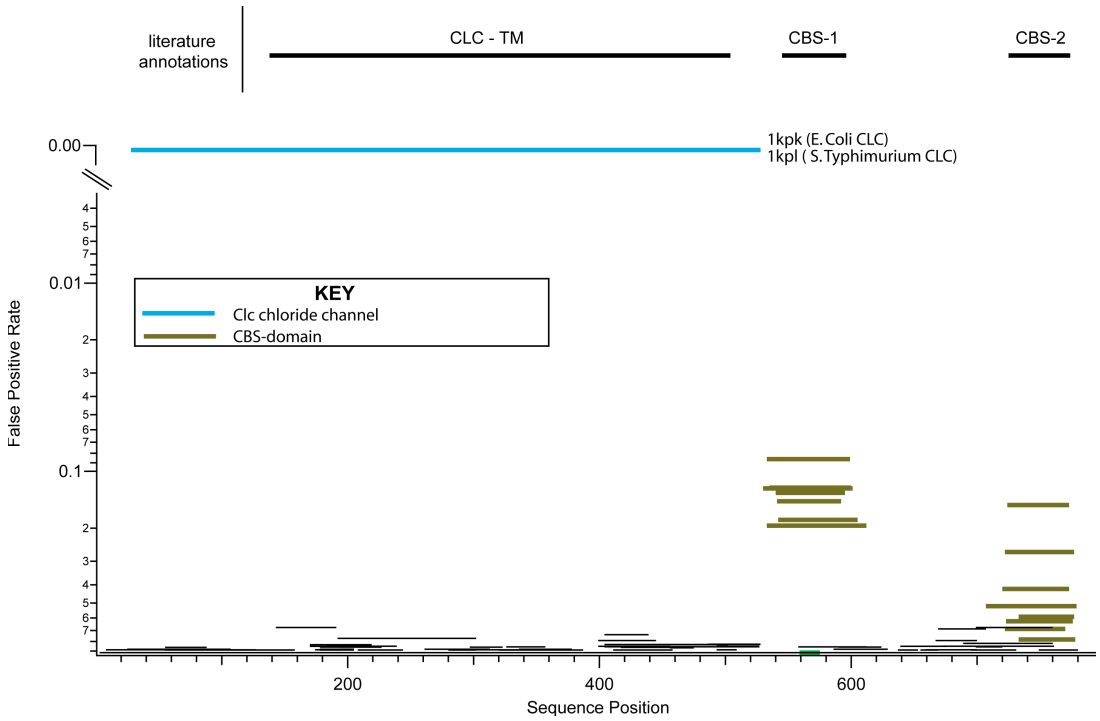

Figure 5. Predicted domains for the Torpedo chloride channel (ClC-0). The results of running the ClC-0 (Genbank/EMBL/DDBJ 544028) channel over the 9,939 Superfamily profile HMMs representing protein domains of known structure. The tabular data used to construct this graph are available online as Table S3. Literature annotations are from Estevez and Jentsch (2002).

calcium channel genes consist of four repeats of the channel motif contained within a single gene. If sodium and calcium channels share a common core domain architecture with potassium and cyclic-nucleotide gated channels, we would expect the solved potassium channel domains to map four times to these genes. This is exactly what we see with a low false positive rate (<0.01) for both the L-type calcium channel gene (Fig. 6) and the human cardiac sodium channel gene (Fig. 7).

It has been shown that internal calcium concentration can affect the inactivation rate of both L-type calcium and human cardiac sodium channels (Peterson et al., 2000; Wingo et al., 2004). In both cases, proposed EF-hand domains have been suggested as being involved, either directly or indirectly, in the observed modulation by calcium (Peterson et al., 2000; Wingo et al., 2004). We see in Figs. 6 and 7 that the proposed EF-hand domains in both channels are found by our methods. While no individual EF-hand domain matches with a false positive rate of less than <0.05, a large number of EF-hand domains match the COOH-terminal region of both channels (Figs. 6 and 7, yellow lines). While a detailed description of the relationship between our confidence in a single strong hits versus multiple moderate hits is beyond the scope of this paper, we point out that identical proteins (>95% identity) are removed from the set of seed sequences used to build the Superfamily database (Gough et al., 2001). Moreover, profiles that regularly generate identical patterns of hits are also removed from the Superfamily database (Gough et al., 2001). There is, therefore, at least some degree of independence between hits. The occurrence of so many moderate hits in the same regions of channel sequence, therefore, supports the assertion for both channels that the COOH-terminal region folds into an EF-hand motif.

Evaluation of Literature-proposed Domains in the BK Channel

In the HCN (Fig. 4), calcium (Fig. 6), and sodium (Fig. 7) channel sequences that we have analyzed, there was strong support for the presence of the S1–S6 "core" regions based on homology to structures of prokaryotic potassium channels. Moreover, in these channels, we were also able to find clear support for proposed cyclic-nucleotide or calcium-binding regulatory domains in the COOH-terminal regions. Everything that we have observed so far is in good agreement with assertions made in the literature.

We now turn to a protein for which the interpretation of literature assignments will not prove as straightforward. The large conductance calcium-activated potassium (BK) channel is gated by both calcium and voltage. A controversy that has surrounded the BK channel concerns the location within the channel sequence of the calcium sensor. The BK channel has a long ~800 amino acid tail after the S6 transmembrane domain.

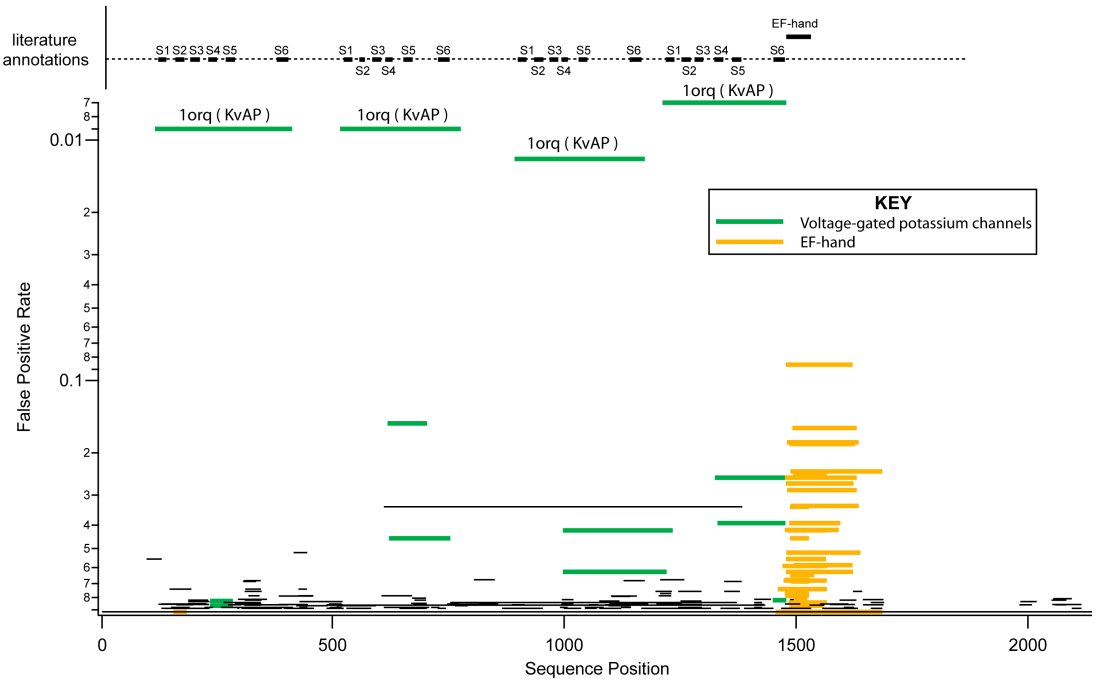

Figure 6. Predicted domains for the mouse voltage-dependent L-type calcium channel. The results of running the L-type calcium channel (Genbank/EMBL/DDBJ 6165982) over the 9,939 Superfamily profile HMMs representing protein domains of known structure. The tabular data used to construct this graph are available online as Table S4. S1–S6 regions map predicted transmembrane domains and are taken from Ma et al. (1995).

The tail of the BK channel is highly conserved between species, for example 95% identical between mouse and human, and does not, using pairwise metrics of sequence similarity, have any immediately obvious homology to any other known protein domain. This domain, highly conserved among, and unique to, BK channels has been the subject of a good deal of interest, much of it regarding whether this domain could harbor the calcium sensor of the channel. A number of different schemes whereby calcium could bind to the BK tail have been proposed (Schreiber and Salkoff, 1997; Jiang et al., 2001; Jiang et al., 2002; Bao et al., 2004). These have included a proposed novel calcium binding domain that has been called the "calcium bowl" (Schreiber and Salkoff, 1997). It has recently been proposed that the calcium bowl region is within a domain that resembles an EF-hand motif (Braun and Sy, 2001; Sheng et al., 2005). The location of the proposed EF-hand domain, based on the published alignment (Sheng et al., 2005), is shown in the top part of Fig. 8.

To what extent is this hypothesis of an EF-hand domain supported by the sequence of the BK channel? Fig. 8 shows the results of running the Mouse BK sequence over the Superfamily database. Hits in which the seed sequence belonged to the EF-hand motif are shown as thick yellow lines. In contrast to the calcium and sodium channels that we examined (Figs. 6 and 7), the EF-hand motif is not found in the post-S6 region of the BK channel with any reasonable false positive rate.

There is evidence beyond sequence analysis to support the EF-hand motif hypothesis. Electrophysiological support for this hypothesis comes from experiments that find changed calcium sensitivity in channels with mutations at residues proposed to correspond to important residues in an EF-hand motif (Braun and Sy, 2001). The degree to which this evidence supports an EF-hand motif in the absence of sequence support is an open question, since it is always a possibility that this region of the channel is in fact an EF-hand domain with a sequence that does not closely resemble other EF-hand domains. We have seen that often domains can belong to the same superfamily yet share little by way of sequence similarity (Fig. 2 A). We point out, however, that there are a large number of different structural models that could explain a given set of biochemical data. This is the primary difficulty in making structural arguments based on mutagenesis. The BK channel could very well have a calcium-sensing fold with a novel structure that could, nonetheless, be consistent with experimental data that appears to support an EF-hand model. This is especially true given that electrophysiological measurements can only report the apparent affinity of ligand for a channel. Mutations that change the underlying energetics of channel gating can cause a shift in the apparent affinity of a ligand without necessarily being near the actual binding site of the ligand. The absence of strong sequence support for proposed domains, therefore, renders

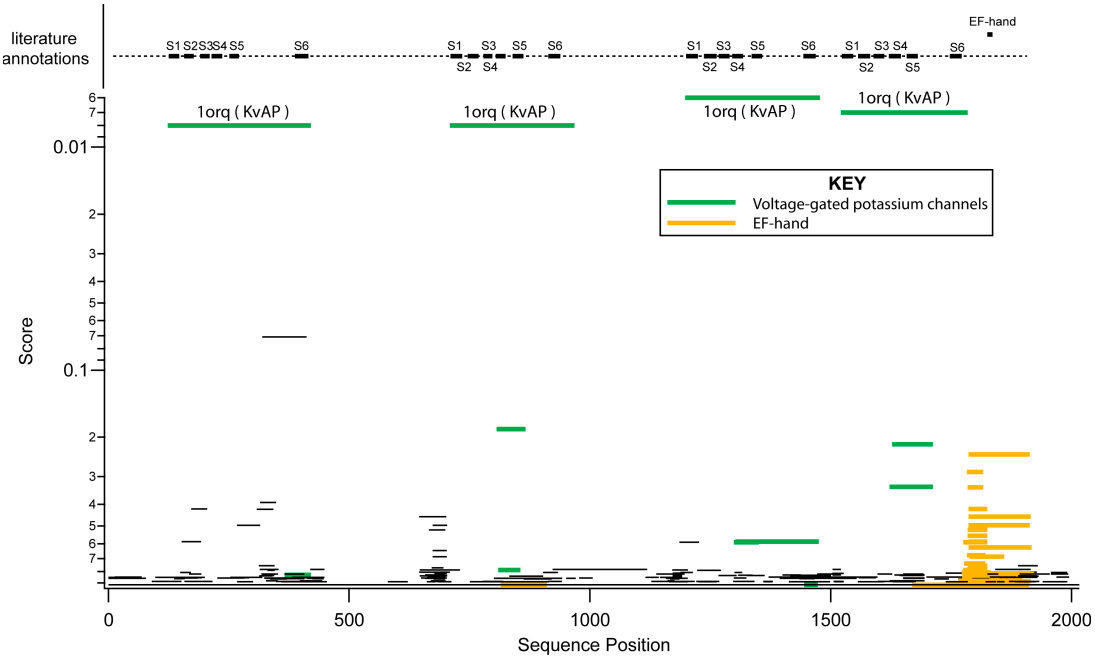

Figure 7. Predicted domains for the human cardiac sodium channel. The results of running the human cardiac sodium channel (Genbank/DMBL/DDBJ 184039) over the 9,939 Superfamily profile HMMs representing protein domains of known structure. The tabular data used to construct this graph are available online as Table S5. S1–S6 regions map predicted transmembrane domains and are taken from Gellens et al. (1992).

interpretation of biochemical and electrophysiological data that much more difficult.

The difficulty of interpreting electrophysiological data in the absence of compelling sequence support can be further illustrated by considering what is essentially a competing hypothesis about the structure of the BK channel. Based on the sensitivity of the channel to serine proteinase inhibitors, and the results of a sequence analysis, it has been proposed that the COOH terminus of the BK channel "structurally resembles serine proteinases" (Moss et al., 1996a; Moss et al., 1996b). The top part of Fig. 8 shows this proposed "serine proteinase-like" domain mapped to the channel sequence based on the published alignment (Moss et al., 1996a). We see that this prediction overlaps the predicted EF-hand domain. The purple lines in the bottom part of Fig. 8 show hits against profiles whose seed sequence belonged to the "trypsin-like serine proteinases" superfamily. As was the case for EF-hands, we see that assignments to this SCOP superfamily do not occur at any reasonable false positive rate (see Table S6). And, yet, it seems inarguable that molecules that inhibit serine proteinases affect the channel (Moss et al., 1996a,b; Favre et al., 2000). Unless this region of the channel can adopt radically different conformations, a possibility that seems unlikely, the EF-hand hypothesis and the serine proteinase-like domain hypothesis are mutually exclusive, despite the existence of supporting biochemical evidence for both hypotheses. The absence of compelling sequence support for either hypothesis makes it difficult to choose between them.

The EF-hand and "calcium bowl" hypotheses have not been the only proposed mechanisms whereby the COOH terminus of the BK channel can sense calcium. In 2001, the MacKinnon lab solved a crystal structure of the COOH-terminal of an *Escherichia coli* potassium channel (Jiang et al., 2001). This region of the channel formed a common structure called a Rossman fold. On the basis of a recursive profile search of Genbank/EMBL/DDBJ, it was proposed that this domain, dubbed RCK or "regulator of potassium conductance" was also present in the BK channel. The position of this RCK1 domain, based on the published alignment, is shown in Fig.8. We see that, in fact, the *E. coli* RCK structure does match the BK channel with a good false positive rate of ~0.03. The assertion that these regions of the two channels share a Rossman fold is therefore highly reasonable. In 2002, however, the MacKinnon lab published a structure of an MthK potassium channel (Jiang et al., 2002). The structure of this channel included a "gating ring" consisting of eight RCK domains, which appeared to be in a position to coordinate calcium and "perform mechanical work to open the pore." Based on sequence analysis, and the fact that potassium channels are tetramers, while the MthK crystal structure showed eight RCK domains apparently coordinating calcium, it was suggested that a second RCK domain existed in the BK channel. Although the exact position of the second

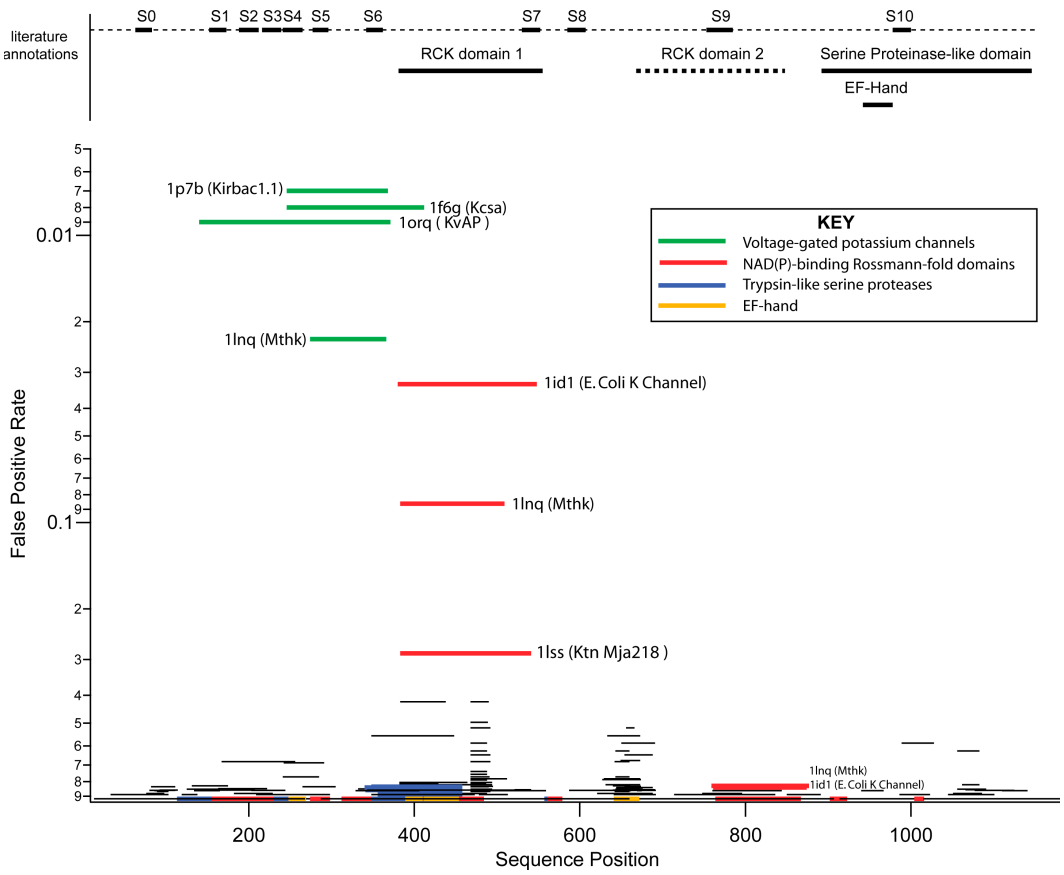

Figure 8. Predicted domains for the mouse calcium-activated potassium (BK) channel. The results of running the mouse BK channel (Genbank/DMBL/DDBJ 487796) over the 9,939 Superfamily profile HMMs representing protein domains of known structure. The tabular data used to construct this graph are available online as Table S6. S0–S10 regions map predicted hydrophobic regions and are taken from Wallner et al. (1996) (S0–S4) and Schreiber et al. (1998) (S5–S10).

RCK domain within the channel structure was not indicated in the MthK paper, it presumably occurs soon after the initial RCK domain and is approximately as long. We have indicated this approximate position as dashed lines in Fig. 8, where we see that, in fact, the second Rossman fold maps to the BK channel with a high false positive rate of >75%.

It is intuitively pleasing to think that the MthK and BK channels work in the same way and share a conserved common mechanism of calcium binding. And the fact that both the MthK and the *E. coli* Rossman fold domains map to the second RCK domain, albeit with high false positive rates (Fig. 8), provide some support for the existence of this domain within BK. In addition, there is some supporting electrophysiological evidence suggesting this region of the channel may be important for sensitivity to calcium (Qian et al., 2002). Nonetheless, we have a great deal more confidence in the existence of the first RCK domain in BK than in the second. In the absence of strong sequence-based evidence, electrophysiological and biochemical evidence for the existence of a second RCK domain must be especially compelling.

DISCUSSION

Ion channels are a physiologically crucial set of proteins that, despite the great progress of the last decade, remain difficult to crystallize. We have used bioinformatic techniques to determine the appropriate level of confidence in our knowledge of the structures of channels for which we lack direct X-ray crystal structure data. To a perhaps surprising degree, we find that the handful of structures that are known are broadly applicable to a wide range of channels. In our survey, the "core" conducting regions of eukaryotic potassium (Fig. 3), HCN (Fig. 4), chloride (Fig. 5), calcium (Fig. 6), and sodium (Fig. 7) channels all map with false positive rates <0.01 to their prokaryotic counterparts. In addition, we have a great deal of confidence in the existence of COOH-terminal modulatory domains for these channels.

The central limitation of our approach is that two sequences may have no discernable sequence similarity and yet may share the same fold (Fig. 2 A). We cannot, therefore, say with certainty that an assertion of common structure is false, even if there appears to be little by way of sequence support for that assertion.

One day, for example, we may have a structure of the BK COOH terminus, and it may very well contain some of the domains that appear in Fig. 8 with high false positive rates. If that turns out to be the case, the combination of intuition, biochemistry, and manual sequence analysis employed by good scientists will have trumped the kinds of automated sequence analyses we perform here. Nonetheless, the initial arguments for the presence of the RCK 2, serine proteinase-like, and EF-hand domains in BK were based in part or in whole on sequence, so it is fair to evaluate the strength of that sequence evidence. In the absence of direct structural data, acceptance of the hypotheses that these domains exist in the channel will require a great deal more experimental work than, for example, confirmation of the first RCK domain in the BK channel, which maps to the channel with a much lower false positive rate (Fig. 8).

The false positive rates that we calculate are dependent on a large number of assumptions and heuristics. We assume, for example, that the PDB is large enough to produce stable results. That is, that the probabilities we calculate won't significantly change as more structures are added. We assume that false positive rates generated from all PDB structures are relevant when applied to ion channels, which, of course, are not well represented in the PDB. We assume that our technique based on the Superfamily database is representative of all possible reasonable techniques. If we had used a different methodology, we might have obtained different results. For example, we could have used a protein threading approach (McGuffin et al., 2004) or a profile–profile (Wang and Dunbrack, 2004) approach rather than running a single channel sequence against a profile HMM to classify domains within our proteins. We could have used a protein classification database other than SCOP or a profile database other than Superfamily. We could have used a program other than HMMer or restricted our analysis to only membrane proteins. While the technique we have used here gives reasonable performance given the current state of the art in detecting structure from sequence (Gough et al., 2001; Madera and Gough, 2002; Madera et al., 2004; Wistrand and Sonnhammer, 2005), we can imagine rational approaches to this problem that would use other methods. Despite the inherent assumptions, we argue that our metric of measuring false positive rates is preferable to the alternative of asserting a common fold between a channel of interest and a channel of known structure based primarily on a visual inspection of a multiple sequence alignment with little assessment as to the statistical merit of that assertion.

One day we might have crystal structures for every protein that we care about, and there will be no need for the kind of bioinformatic estimates we have discussed here. In the meantime, by explicitly considering estimates of error rates in assertions of common structure, we can focus our experimental efforts on the most probable structural models for the proteins we study.

We thank Merritt Maduke, Weiyan Li, Andrea Meredith, Jon Sack, and Christina Wilkens for reading this manuscript.

This work was supported by the Mathers Foundation.

Olaf S. Andersen served as editor.

Submitted: 30 September 2005
Accepted: 15 May 2006

REFERENCES

Accardi, A., and C. Miller. 2004. Secondary active transport mediated by a prokaryotic homologue of ClC Cl^- channels. *Nature.* 427:803–807.

Andreeva, A., D. Howorth, S.E. Brenner, T.J. Hubbard, C. Chothia, and A.G. Murzin. 2004. SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.* 32:D226–D229.

Bao, L., C. Kaldany, E.C. Holmstrand, and D.H. Cox. 2004. Mapping the BKCa channel's "Ca^{2+} bowl": side-chains essential for Ca^{2+} sensing. *J. Gen. Physiol.* 123:475–489.

Braun, A.P., and L. Sy. 2001. Contribution of potential EF hand motifs to the calcium-dependent gating of a mouse brain large conductance, calcium-sensitive K(+) channel. *J. Physiol.* 533:681–695.

Brenner, S.E., C. Chothia, and T.J. Hubbard. 1998. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci. USA.* 95:6073–6078.

Brenner, S.E., P. Koehl, and M. Levitt. 2000. The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res.* 28:254–256.

Chandonia, J.M., G. Hon, N.S. Walker, L. Lo Conte, P. Koehl, M. Levitt, and S.E. Brenner. 2004. The ASTRAL compendium in 2004. *Nucleic Acids Res.* 32:D189–D192.

Chandonia, J.M., N.S. Walker, L. Lo Conte, P. Koehl, M. Levitt, and S.E. Brenner. 2002. ASTRAL compendium enhancements. *Nucleic Acids Res.* 30:260–263.

DiFrancesco, D., and P. Tortora. 1991. Direct activation of cardiac pacemaker channels by intracellular cyclic AMP. *Nature.* 351:145–147.

Durbin, R., S.R. Eddy, A. Krogh, and G. Mitchison. 1998. Biological Sequence Analysis: Probablistic Models of Proteins and Nucleic Acids. Cambridge University Press, Cambridge, UK.. 368 pp.

Dutzler, R., E.B. Campbell, M. Cadene, B.T. Chait, and R. MacKinnon. 2002. X-ray structure of a ClC chloride channel at 3.0 A reveals the molecular basis of anion selectivity. *Nature.* 415:287–294.

Dutzler, R., E.B. Campbell, and R. MacKinnon. 2003. Gating the selectivity filter in ClC chloride channels. *Science.* 300:108–112.

Estevez, R., and T.J. Jentsch. 2002. CLC chloride channels: correlating structure with function. *Curr. Opin. Struct. Biol.* 12:531–539.

Favre, I., G.W. Moss, D.P. Goldenberg, J. Otlewski, and E. Moczydlowski. 2000. Structure-activity relationships for the interaction of bovine pancreatic trypsin inhibitor with an intracellular site on a large conductance Ca(2+)-activated K(+) channel. *Biochemistry.* 39:2001–2012.

Gellens, M.E., A.L. George Jr., L.Q. Chen, M. Chahine, R. Horn, R.L. Barchi, and R.G. Kallen. 1992. Primary structure and functional expression of the human cardiac tetrodotoxin-insensitive voltage-dependent sodium channel. *Proc. Natl. Acad. Sci. USA.* 89:554–558.

Gough, J., K. Karplus, R. Hughey, and C. Chothia. 2001. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.* 313:903–919.

Hubbard, T.J., B. Ailey, S.E. Brenner, A.G. Murzin, and C. Chothia. 1998. SCOP, structural classification of proteins database: applications to evaluation of the effectiveness of sequence alignment methods and statistics of protein structural data. *Acta Crystallogr. D Biol. Crystallogr.* 54:1147–1154.

Hubbard, T.J., B. Ailey, S.E. Brenner, A.G. Murzin, and C. Chothia. 1999. SCOP: a structural classification of proteins database. *Nucleic Acids Res.* 27:254–256.

Hubbard, T.J., A.G. Murzin, S.E. Brenner, and C. Chothia. 1997. SCOP: a structural classification of proteins database. *Nucleic Acids Res.* 25:236–239.

Jiang, Y., A. Lee, J. Chen, M. Cadene, B.T. Chait, and R. MacKinnon. 2002. Crystal structure and mechanism of a calcium-gated potassium channel. *Nature.* 417:515–522.

Jiang, Y., A. Lee, J. Chen, V. Ruta, M. Cadene, B.T. Chait, and R. MacKinnon. 2003. X-ray structure of a voltage-dependent K+ channel. *Nature.* 423:33–41.

Jiang, Y., A. Pico, M. Cadene, B.T. Chait, and R. MacKinnon. 2001. Structure of the RCK domain from the *E. coli* K+ channel and demonstration of its presence in the human BK channel. *Neuron.* 29:593–601.

Kreusch, A., P.J. Pfaffinger, C.F. Stevens, and S. Choe. 1998. Crystal structure of the tetramerization domain of the Shaker potassium channel. *Nature.* 392:945–948.

Lo Conte, L., B. Ailey, T.J. Hubbard, S.E. Brenner, A.G. Murzin, and C. Chothia. 2000. SCOP: a structural classification of proteins database. *Nucleic Acids Res.* 28:257–259.

Lo Conte, L., S.E. Brenner, T.J. Hubbard, C. Chothia, and A.G. Murzin. 2002. SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.* 30:264–267.

Long, S.B., E.B. Campbell, and R. Mackinnon. 2005a. Crystal structure of a mammalian voltage-dependent Shaker family K+ channel. *Science.* 309:897–903.

Long, S.B., E.B. Campbell, and R. Mackinnon. 2005b. Voltage sensor of Kv1.2: structural basis of electromechanical coupling. *Science.* 309:903–908.

Ma, Y., E. Kobrinsky, and A.R. Marks. 1995. Cloning and expression of a novel truncated calcium channel from non-excitable cells. *J. Biol. Chem.* 270:483–493.

Madera, M., and J. Gough. 2002. A comparison of profile hidden Markov model procedures for remote homology detection. *Nucleic Acids Res.* 30:4321–4328.

Madera, M., C. Vogel, S.K. Kummerfeld, C. Chothia, and J. Gough. 2004. The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Res.* 32:D235–D239.

McGuffin, L.J., S. Street, S.A. Sorensen, and D.T. Jones. 2004. The genomic threading database. *Bioinformatics.* 20:131–132.

McKinnon, D. 1989. Isolation of a cDNA clone coding for a putative second potassium channel indicates the existence of a gene family. *J. Biol. Chem.* 264:8230–8236.

Minor, D.L., Y.F. Lin, B.C. Mobley, A. Avelar, Y.N. Jan, L.Y. Jan, and J.M. Berger. 2000. The polar T1 interface is linked to conformational changes that open the voltage-gated potassium channel. *Cell.* 102:657–670.

Moss, G.W., J. Marshall, and E. Moczydlowski. 1996a. Hypothesis for a serine proteinase-like domain at the COOH terminus of Slowpoke calcium-activated potassium channels. *J. Gen. Physiol.* 108:473–484.

Moss, G.W., J. Marshall, M. Morabito, J.R. Howe, and E. Moczydlowski. 1996b. An evolutionarily conserved binding site for serine proteinase inhibitors in large conductance calcium-activated potassium channels. *Biochemistry.* 35:16024–16035.

Murzin, A.G., S.E. Brenner, T. Hubbard, and C. Chothia. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247:536–540.

Peterson, B.Z., J.S. Lee, J.G. Mulle, Y. Wang, M. de Leon, and D.T. Yue. 2000. Critical determinants of Ca(2+)-dependent inactivation within an EF-hand motif of L-type Ca(2+) channels. *Biophys. J.* 78:1906–1920.

Qian, X., C.M. Nimigean, X. Niu, B.L. Moss, and K.L. Magleby. 2002. Slo1 tail domains, but not the Ca^{2+} bowl, are required for the β1 subunit to increase the apparent Ca2+ sensitivity of BK channels. *J. Gen. Physiol.* 120:829–843.

Rost, B. 1999. Twilight zone of protein sequence alignments. *Protein Eng.* 12:85–94.

Rychlewski, L., L. Jaroszewski, W. Li, and A. Godzik. 2000. Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci.* 9:232–241.

Santoro, B., D.T. Liu, H. Yao, D. Bartsch, E.R. Kandel, S.A. Siegelbaum, and G.R. Tibbs. 1998. Identification of a gene encoding a hyperpolarization-activated pacemaker channel of brain. *Cell.* 93:717–729.

Schreiber, M., and L. Salkoff. 1997. A novel calcium-sensing domain in the BK channel. *Biophys. J.* 73:1355–1363.

Schreiber, M., A. Wei, A. Yuan, J. Gaut, M. Saito, and L. Salkoff. 1998. Slo3, a novel pH-sensitive K+ channel from mammalian spermatocytes. *J. Biol. Chem.* 273:3509–3516.

Scott, J.W., S.A. Hawley, K.A. Green, M. Anis, G. Stewart, G.A. Scullion, D.G. Norman, and D.G. Hardie. 2004. CBS domains form energy-sensing modules whose binding of adenosine ligands is disrupted by disease mutations. *J. Clin. Invest.* 113:274–284.

Sheng, J.Z., A.M. Weljie, L. Sy, S. Ling, H.J. Vogel, and A.P. Braun. 2005. Homology modeling identifies C-terminal residues that contribute to the Ca2+ sensitivity of a BKCa channel. *Biophys J.* 89:3079–3092.

Wallner, M., P. Meera, and L. Toro. 1996. Determinant for beta-subunit regulation in high-conductance voltage-activated and Ca(2+)-sensitive K+ channels: an additional transmembrane region at the N terminus. *Proc. Natl. Acad. Sci. USA.* 93:14922–14927.

Wang, G., and R.L. Dunbrack Jr. 2004. Scoring profile-to-profile sequence alignments. *Protein Sci.* 13:1612–1626.

Wingo, T.L., V.N. Shah, M.E. Anderson, T.P. Lybrand, W.J. Chazin, and J.R. Balser. 2004. An EF-hand in the sodium channel couples intracellular calcium to cardiac excitability. *Nat. Struct. Mol. Biol.* 11:219–225.

Wistrand, M., and E.L. Sonnhammer. 2005. Improved profile HMM performance by assessment of critical algorithmic features in SAM and HMMER. *BMC Bioinformatics.* 6:99.

Zagotta, W.N., N.B. Olivier, K.D. Black, E.C. Young, R. Olson, and E. Gouaux. 2003. Structural basis for modulation and agonist specificity of HCN pacemaker channels. *Nature.* 425:200–205.