doi:10.1016/j.jmb.2010.08.032 J. Mol. Biol. (2010) 403, 480–493

Contents lists available at www.sciencedirect.com

Journal of Molecular Biology

journal homepage: http://ees.elsevier.com.jmb

The Evolution and Structure Prediction of Coiled Coils across All Genomes

Owen J. L. Rackham[1,2], Martin Madera[1], Craig T. Armstrong[3], Thomas L. Vincent[2,3], Derek N. Woolfson[3,4] and Julian Gough[1]*

[1]*Department of Computer Science, University of Bristol, Bristol BS8 1UB, UK*
[2]*Bristol Centre for Complexity Sciences, University of Bristol, Bristol BS8 1TR, UK*
[3]*School of Chemistry, University of Bristol, Bristol BS8 1TS, UK*
[4]*Department of Biochemistry, University of Bristol, Bristol BS8 1TD, UK*

Received 22 February 2010; received in revised form 6 August 2010; accepted 17 August 2010
Available online 9 September 2010

Edited by M. Sternberg

Keywords:
protein evolution;
coiled coil;
hidden Markov model;
protein structure;
SUPERFAMILY

Coiled coils are α-helical interactions found in many natural proteins. Various sequence-based coiled-coil predictors are available, but key issues remain: oligomeric state and protein–protein interface prediction and extension to all genomes. We present SpiriCoil (http://supfam.org/SUPERFAMILY/spiricoil), which is based on a novel approach to the coiled-coil prediction problem for coiled coils that fall into known superfamilies: hundreds of hidden Markov models representing coiled-coil-containing domain families. Using whole domains gives the advantage that sequences flanking the coiled coils help. SpiriCoil performs at least as well as existing methods at detecting coiled coils and significantly advances the state of the art for oligomer state prediction. SpiriCoil has been run on over 16 million sequences, including all completely sequenced genomes (more than 1200), and a resulting Web interface supplies data downloads, alignments, scores, oligomeric state classifications, three-dimensional homology models and visualisation. This has allowed, for the first time, a genomewide analysis of coiled-coil evolution. We found that coiled coils have arisen independently *de novo* well over a hundred times, and these are observed in 16 different oligomeric states. Coiled coils in almost all oligomeric states were present in the last universal common ancestor of life. The vast majority of occasions that individual coiled coils have arisen *de novo* were before the last universal common ancestor of life; we do, however, observe scattered instances throughout subsequent evolutionary history, mostly in the formation of the eukaryote superkingdom. Coiled coils do not change their oligomeric state over evolution and did not evolve from the rearrangement of existing helices in proteins; coiled coils were forged in unison with the fold of the whole protein.

*Corresponding author. E-mail address: gough@cs.bris.ac.uk.

Abbreviations used: KIH, knobs-into-holes; HMM, hidden Markov model; PDB, Protein Data Bank; 3D, three-dimensional; LUCA, last universal common ancestor of life.

Introduction

The ability to rapidly sequence DNA has led to an explosion in the amount of sequence data available for partial and whole genomes. As a result, the field of genomics, with its various subfields, has

emerged. These now allow systematic analyses of many aspects of biology. Of particular interest to us is the evolution of protein sequences and how this relates to the evolution of protein structures. Here we present a new approach to predicting one type of protein secondary structure interaction, namely the α-helical coiled coil.[1]

Unlike most protein structures, α-helical coiled coils are relatively straightforward. They comprise two or more α-helices wrapped around each other to form compact helical bundles or rope-like extended structures (Fig. 1a and b). The structures range from simple dimers through pentamers, to more complex assemblies involving many helices and even bundles of bundles.[2–4] In these ways, coiled coils cement helix–helix interactions within all classes of protein structure (globular, fibrous and membrane), although they are most often linked with protein–protein interactions.[1,5,6] Notably, coiled coils play key roles in directing and controlling protein–protein associations involved in transcription, membrane fusion, muscle structure and function, the cytoskeleton, the extracellular matrix, cell division, the chemotaxis machinery, host–pathogen interactions, and many other events within and outside the cell.

Given the proliferation and the structural and functional diversity of coiled coils, it is somewhat surprising that the sequences underlying them are, at first sight, quite simple: typically, coiled-coil sequences display a repeat pattern of hydrophobic (H) and polar (P) residues, HPPHPPP, which usually extends for 20 or more residues but can be shorter or can span many hundreds of amino acids. This pattern is also called the *heptad repeat* and is often labelled *abcdefg*, with the *a* and *d* positions corresponding to the H type. Thus, hydrophobic residues are alternately spaced three and four residues apart. This average spacing of 3.5 is close to the 3.6 residues/turn of a regular α-helix. As a result, when configured into α-helices, coiled-coil sequences give amphipathic structures with distinct hydrophobic and polar faces (Fig. 1c). The association of two or more such helices through their hydrophobic sites drives coiled-coil assembly. However, because 3.5 is smaller than 3.6, to maintain hydrophobic contacts, the helices wrap or supercoil around one another with a left-handed twist (Fig. 1a and b). Furthermore, the hydrophobic residues at *a* and *d* from one helix form 'knobs' that pack into 'holes' formed by a combination of residues at *g*, *a*, *d* and *e* on partnering helices (Fig. 1d). This so-called 'knobs-into-holes' (KIH) packing, first posited by Crick, represents a direct and rational link between coiled-coil sequence and structure, which is one of the most transparent in the protein world.[7] Note that variations on the heptad pattern are known[8,9] and are increasingly becoming apparent;[6] however, these so-called noncanonical coiled coils are also based on repeats with H-type and P-type residues

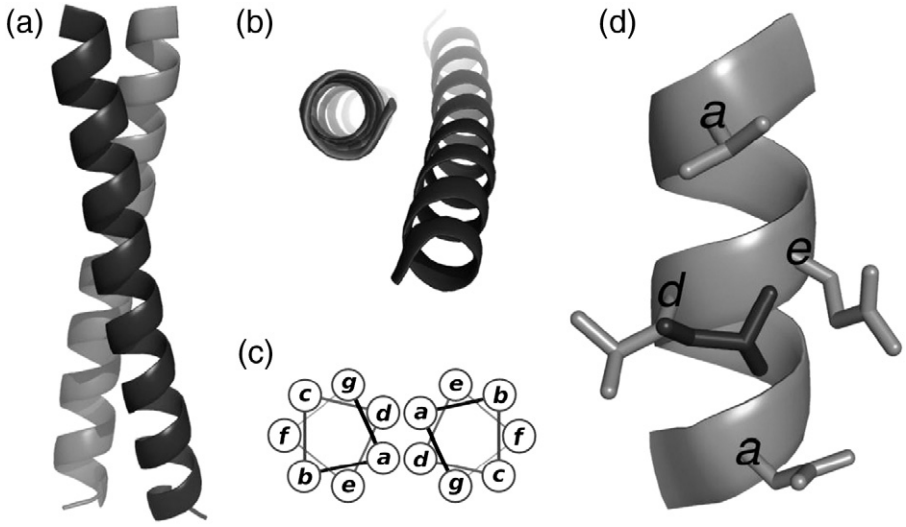

Fig. 1. The coiled-coil structure and sequence. (a) Ribbon diagram showing the two helices of a parallel dimeric coiled coil (PDB entry 2ZTA). (b) Orthogonal view to (a) looking down the coiled-coil axis. The helices pack slightly off the alignment, allowing their hydrophobic stripes to mesh as described in the text. (c) The heptad repeat *abcdefg* typical of coiled-coil sequences spun out on a helical wheel, which captures in cartoon form the helices shown in (b). Hydrophobic residues tend to occupy the *a* and *d* sites; as (c) shows, these fall on the same face of each helix and thus lead to helix association. (d) A KIH interaction from the 2ZTA structure. The rearmost helix is shown in light grey, with side chains at consecutive *a*, *d*, *e* and *a* sites drawn as sticks. These form a "hole" that accepts a "knob" at a *d* site (dark grey) from the partnering helix (data not shown for clarity).

spaced at combinations of three and four residues apart, leading to slight variations in KIH packing and different degrees of helical supercoiling.[6,9]

These clear links between sequence and structure have led to a number of structure prediction algorithms that take only protein sequence as their input. The first of these, COILS from Lupas *et al.*,[3] is based on the concept of amino acid profiles (or position-specific scoring matrix) from Parry[10] and has recently been updated to PCOILS.[11] Other coiled-coil prediction methods have employed simple masks for HP patterns (e.g., COILER)[12] or pairwise information between residues of the heptad repeat (e.g., Paircoil and Paircoil2 from McDonnell *et al.*[13] and Berger *et al.*[14]). A key development was the application of hidden Markov models (HMMs) to coiled-coil prediction, which led to MARCOIL from Delorenzi and Speed[15] and, most recently, to CCHMM_PROF from Bartoli *et al.*[16] COILS is probably still the most widely used method, although it is outperformed by HMM-based methods.[16] In addition, MultiCoil is available and able to predict oligomer state from sequence,[12,17] building on the Paircoil explained above. However, it has limited scope as it is only a two-state (dimer or trimer) predictor, whereas it is becoming increasingly apparent that many more oligomer states, topologies and complex assemblies are possible within the coiled-coil framework.[2–4]

The key contemporary issues in coiled-coil prediction are as follows: (1) reliable identification of coiled-coil regions within protein sequences; (2) multistate prediction of oligomer state based on this information; and (3) extension of these methods to prediction at the level of systems and whole genomes.

Several groups have performed coiled-coil predictions at the genome level to deliver what might be termed 'coilomes':[18] Newman *et al.*[19] described an analysis of *Saccharomyces cerevisiae*, which they followed up with a yeast two-hybrid analysis to report >200 interactions among >150 proteins, whereas Rose *et al.* developed the ARABI-COIL database of long coiled-coil proteins predicted for *Arabidopsis thaliana*[20] and explored similar predictions across 22 proteomes.[21] These three studies used MultiCoil as predictor and reported coilomes of ~2–8% of the parent proteomes. In another analysis of 29 complete proteomes, Liu and Rost showed that COILS reports ~10% coiled coils in eukaryotes and 4–5% coiled coils in prokaryotes and Archaea.[22] Most recently, Barbara *et al.* again combined coiled-coil predictions and experiments to further define the *S. cerevisiae* coilome.[18] In this case, using COILS and Paircoil2, they suggested that up to 20% of the proteome contains coiled coils. Ideally, new genomewide studies of potential coilomes would also consider predictions for the broader gamete of possible coiled-coil oligomeric states, cover all the currently completed genomes and be regularly updated.

Here we address these issues via a structurally informed homology-based prediction method, which not only identifies coiled coils in protein sequences but also includes the prediction of all oligomeric states. It is fundamentally different from the existing sequence-based approaches in that it does not rely on a single model, profile or statistical scoring for coiled coils. Instead, we employ a large library of HMMs representing protein domains that are known to contain coiled coils. It is an extension of SUPERFAMILY,[23,24] which currently does not include coiled-coil classes. In this approach, the known structures of protein domains and their related sequences are grouped, and this information is used to build multiple HMMs. The key issue in the SUPERFAMILY approach is: What should be included as coiled-coil-positive structures? As others have posited,[11] we used the intersection of the coiled-coil class from SCOP (*S*tructural *C*lassification *o*f *P*roteins)[25,26] (the basis for SUPERFAMILY) and an analysis of the Protein Data Bank (PDB) using SOCKET. The SOCKET results were curated to identify the domains in other classes in SCOP that contain coiled coils. SOCKET is an algorithm[27] that identifies KIH packing from coordinates of known protein structures. Recently, we have used SOCKET to create a database of coiled coils organised in a 'periodic table of coiled-coil protein structures'.[4,28] Thus, with a solid starting point of known coiled-coil structures, we generated new HMMs for coiled-coil-containing proteins. These were used to identify domains in protein sequences that contain coiled coils in superfamilies of known structure, which is not an explicit requirement of other methods. This requirement, unlike other methods, means that the most closely related experimentally solved structure can also be predicted. This locates the coiled-coil region and allows three-dimensional (3D) homology modeling using the known structure as template. Other advantages of going via natural sequences and structures are that all oligomeric states are predicted and extra information on the sequence flanking coiled-coil regions strengthens the predictions.

This approach has been used to annotate all (~1200) completely sequenced genomes and other data sets such as UniProt, viruses and structural genomics targets. The software, SCOP annotation, genome assignments, oligomeric states and 3D models have all been made available via SpiriCoil†, a Web resource for coiled coils in genomes. SpiriCoil includes an annotation pipeline ensuring regular updates as new genomes are sequenced. With periodic HMM updates as new structures are deposited in the PDB and classified in SCOP, the quality and coverage of the predictions will

† http://supfam.org/SUPERFAMILY/spiricoil

progressively improve over time. Importantly, the data in SpiriCoil allow us to begin to understand the evolution of coiled coils across all species in a way that hitherto has not been possible by examining sequences and structures without knowledge of their genomic context.

Results

The principle of SpiriCoil is to treat coiled-coil prediction as a standard homology recognition problem. This does raise an operational issue, however: as coiled-coil sequences share the heptad (or related repeats), they have low complexity and are more likely to be similar even if they are not homologous in the true sense of the term. To address this, we adopted the SUPERFAMILY approach, which has been proven and recently extended to another 'difficult class' of protein structures, namely membrane proteins.[29] Two major steps were required to implement this approach: firstly, all of the domain superfamilies and families in SCOP that contain coiled-coil regions had to be identified along with their oligomeric state; and, secondly, HMMs had to be constructed for those domains not already in SUPERFAMILY and added to the SUPERFAMILY pipeline.

Coiled coils in SCOP

The SCOP database is divided into eight major classes of interest, and one of these (dedicated to coiled-coil domains) was previously not included in SUPERFAMILY. We had to construct new HMMs for all of the domains belonging to the 55 superfamilies in this class. In addition, for SpiriCoil to capture all domains, we also needed to identify all of the coiled-coil domains in other classes (for which we could use existing HMMs). For oligomeric state prediction, we needed to classify the oligomeric states for all of these identified coiled-coil domains. We employed SOCKET,[27] as well as the resulting periodic table of coiled-coil protein structures,[4] to provide a hand-verified set of coiled coils and their corresponding oligomeric state. Specifically, SOCKET was run on the entire PDB (filtered to a 95% sequence identity) as SCOP version 1.73. Both the resulting set and the periodic table were mapped onto the corresponding SCOP family. SOCKET is an automatic procedure; therefore, classification for each family was verified by hand. In addition to the 55 superfamilies in the SCOP coiled-coil class, we found 64 superfamilies in other classes with coiled coils within the domain. These covered 16 different oligomeric states. Of the total 119 superfamilies, 19 did not have total participation of families within their superfamily; thus, our final representation of coiled coils in SCOP had 100 superfamilies and 25 families (contained in 19 superfamilies).

We classified some coiled coils found by SOCKET as 'fuzzy'. On downstream analysis, we still predicted them but labelled them separately from the more clearly defined cases. The fuzzy coiled-coil set in SCOP included 29 superfamilies and 26 families with incomplete participation; each had some members predicted as coiled coils by SOCKET and other members not predicted as coiled coils, usually due to minor structural changes and shifting helices that cause the KIH packing to be less marked. These types, as well as the rest, are listed in Supplementary Information but include four-helix bundles, barrel-like membrane proteins described by us previously[2] and other proteins where many, often short, helices come together with less extensive KIH interactions. In addition to regular and fuzzy coiled coils, we have a third category—'short coiled coils'—with less than 14 residues per helix. There were 69 short coiled-coil-containing superfamilies made up of 51 superfamilies and 28 families (contained in 18 superfamilies). The coiled coils in SCOP can be found in Supplementary Information and are summarised in Table 1.

Testing and benchmarking SpiriCoil predictions

Using the test set described in Materials and Methods, we compared SpiriCoil predictions on known 3D structures with those made by MARCOIL and Paircoil2 (Fig. 2).[13,15] In this receiver operating characteristic plot, false positives and true positives are plotted cumulatively on the x and y axes, respectively. From the origin, the curves are plotted in descending order of confidence provided by each method. These results show that, except for extremes of high error rate (i.e., on the right-hand

Table 1. Number of coiled-coil-containing superfamilies

	Regular	Fuzzy	Short	Total
Superfamilies	119	52	69	280
Complete superfamilies (families)	**100** (106)	**29** (30)	**51** (53)	**180** (189)
Incomplete superfamilies (families)	19 (**25**)	23 (**26**)	18 (**28**)	60 (**79**)

The 'fuzzy' and 'short' classes of coiled coils are described fully in the text. If all families within a superfamily are coiled-coil-containing, it is counted as a 'complete' superfamily and listed in the second row; the total number of families in all those superfamilies is shown in parentheses. If a superfamily comprises some coiled-coil-containing families and some non-coiled-coil-containing families, it is counted as an 'incomplete' superfamily and listed in the third row; the total number of coiled-coil-containing families within those superfamilies is shown in parentheses. Thus, shown in boldface are the coiled-coil-containing entities (e.g., 100 superfamilies plus 25 families in the first column).

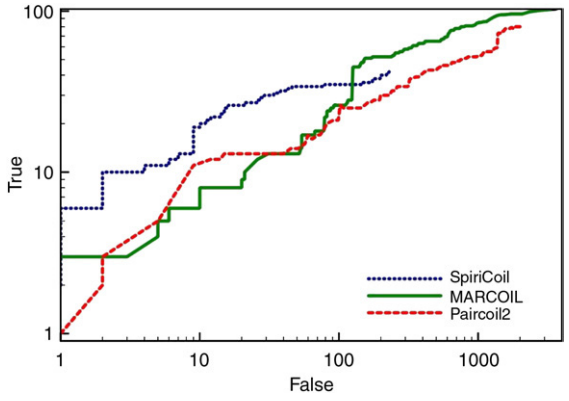

Fig. 2. Coiled-coil detection benchmarking results on known structures. The predictions made by each of the three methods are plotted from the origin in descending order of confidence, counting true positives and false positives cumulatively on the *y* and *x* axes using a logarithmic scale. The left half of the graph is the region of interest to biologists; predictions at very high error rate (low confidence) are of little use.

side where the methods make over three times as many false predictions as true predictions), SpiriCoil outperforms existing techniques. However, a striking and worrisome feature of the plot is that the absolute performance of all methods at first appears to be poor in absolute terms. For SpiriCoil, the first 19 true predictions were made at a rate of 10 false positives; ultimately, only 43 of the possible 103 coiled coils were identified, at which point 237 coiled coils not identified by SOCKET had also been predicted. To understand this, we undertook the analysis of false negatives and false positives below, which shows that although the benchmark gives a good independent measure of relative performance, it is a poor estimate of absolute performance.

False negatives

So what does SpiriCoil miss, and why? We analysed the false negatives (i.e., the 60 coiled coils that were missed) and broke them down into the categories summarised in Table 2. This revealed that over half (33 of 60) were actually identified by SpiriCoil as containing 'fuzzy' or 'short' coiled coils and hence are not completely false negatives. These two categories contain proteins for which SOCKET detects coiled coils intermittently. The remaining 27 proteins appear to contain genuine coiled coils missed by our technique. We examined the superfamilies/families to which these were assigned in order to ascertain why SpiriCoil did not detect them. In some of these superfamilies/families, we had already observed coiled coils identified by SOCKET, but because the majority of the family members did not contain coiled coils, we did not include that family in SpiriCoil. This is a limitation of our approach: we always label a whole family with the same classification, which is usually a reasonable approximation. In rare cases, however, some structural characteristics are different for individuals despite coming from a common evolutionary ancestor. The results from SOCKET are shown in Table 3, and the last column shows the proportion of the members of a superfamily in which SOCKET detects a coiled coil in SCOP

Table 2. Breakdown of SpiriCoil false negatives from the benchmark in Fig. 2

Category	Frequency
Confirmed false negatives	27
Marginal coiled coils	33
Fuzzy	28
Short	5
Total	60

These include confirmed false negatives or marginal coiled coils. Marginal coiled coils are sequences that were predicted as coiled-coil-containing by SpiriCoil but are expected to be members of families/superfamilies that are annotated as containing short coiled coils or as having fuzzy membership, as described in the text.

Table 3. SUPERFAMILY annotation of confirmed false-positive superfamilies

Superfamily	Frequency	CC *versus* non-CC in training
Winged-helix DNA-binding domain	3	3 *versus* 206
Six-hairpin glycosidases	1	4 *versus* 37
Multiheme cytochromes	3	1 *versus* 68
cAMP-binding-domain-like	2	7 *versus* 43
NAD(P)-binding Rossman fold domains	1	1 *versus* 401
CheY-like	1	2 *versus* 105
Restriction-endonuclease-like	1	No true coiled coils
Carbamate-kinase-like	1	1 *versus* 17
UBC-like	1	No true coiled coils
Nucleotide cyclase	1	No true coiled coils
Protein-kinase-like (PK-like)	5	4 *versus* 205
Docking domain A of the erythromycin polyketide synthase	1	No true coiled coils
HBS1-like domain	1	No true coiled coils
SipA N-terminal-domain-like	1	No true coiled coils
NE0471 N-terminal-domain-like	1	No true coiled coils
DEATH domain	1	No true coiled coils
NTF2-like	1	No true coiled coils
ABC transporter transmembrane region	1	No true coiled coils

Some of the 27 superfamilies where false negatives were found in the benchmarking are in superfamilies that, during the annotation, did not contain any coiled coils or contained coiled coils in insufficient number to warrant annotating the superfamily as coiled-coil-containing. The last column shows how many coiled coils were already predicted by SOCKET in those superfamilies.

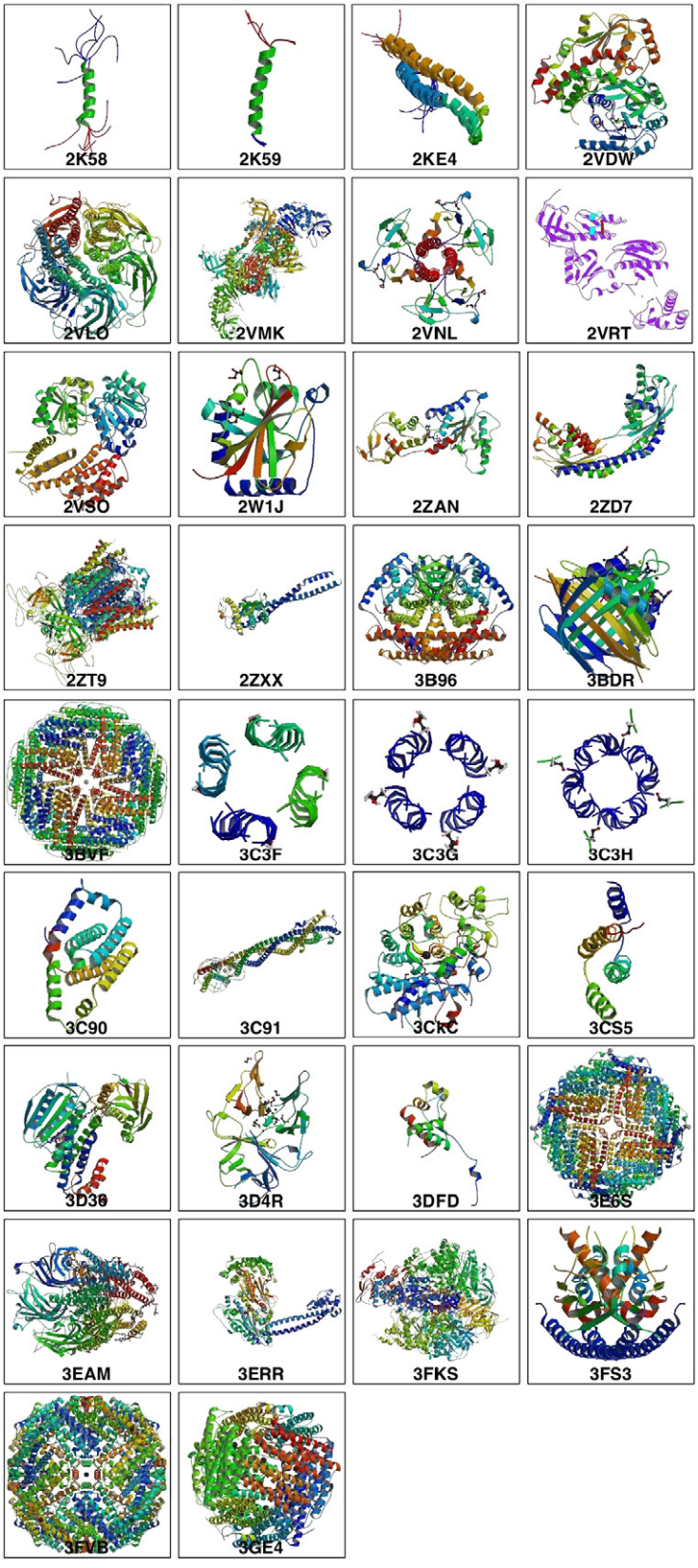

Fig. 3. Structures of SpiriCoil false positives from the coiled-coil detection benchmark. In most cases, it can be seen that these are, in fact, coiled coils and therefore are not false positives. See Results for details. Images were taken from the PDB Web site (http://www.pdb.org).

version 1.73. As more 3D protein structures are solved, the quality and resolution of SpiriCoil will improve (e.g., new structural information may lead SCOP to split the existing classifications, creating new families, consistent with the presence of coiled coils). It is also possible that new coiled-coil structures will emerge in an existing family, enabling us to refine our original annotation and thus shifting our assessment of that family to predominantly coiled-coil-containing.

False positives

We analysed the proteins predicted by SpiriCoil but for which SOCKET had not identified a coiled coil. These amounted to 232 of the 3802 protein structures spread over 92 superfamilies/families. These are cases where SpiriCoil predicted a domain belonging to a family/superfamily expected to contain a coiled coil but where SOCKET did not identify a coiled coil. Of these proteins, 24 were resolved using NMR, which SOCKET can have trouble dealing with and so should be disregarded. Of the remainder, only 34 predictions (15%) had a high confidence value (*E*-value better than 0.01) (Fig. 3). Of these, 3 were NMR structures, 8 did contain coiled coils but issues with PDB files caused SOCKET to fail, 2 were detected by SOCKET but below the 14-residue threshold we had employed and 19 were declared to be coiled coils from visual inspection despite SOCKET not detecting them. This leaves only two of the high-confidence predictions as definite false positives. The false positives with low confidence scores were not analysed, but HMM *E*-values are known to be reliable, and one would expect the ratio of truly false positives to rise rapidly as confidence decreases.

In summary and taking into account these interpretations of the SOCKET results, SpiriCoil found 61 of 88 possible coiled-coil proteins plus two false positives at an *E*-value threshold of 0.01. We did not carry out an in-depth analysis of false positives and false negatives for the other prediction methods, but we may conclude that the absolute performance for SpiriCoil is far better than it would first appear from Fig. 2 based on raw SOCKET results. Taken in isolation, the benchmark may be a poor indication of the absolute performance of the methods, but it remains a fair comparison of relative performance. We can summarise that SpiriCoil also performs slightly better than Paircoil2 and MARCOIL, except in the region of unreasonably high error rate. SpiriCoil has one limitation—it tars all domains that are members of the same SCOP family with the same brush. This is not a serious limitation since, except in the case of fuzzy and short coiled coils, a family is almost always homogeneous.

Oligomeric state prediction with SpiriCoil

Moving on from coiled-coil detection *per se*, a contemporary issue in coiled-coil prediction is the classification of oligomeric state: dimer, trimer, tetramer and so on. One method, namely MultiCoil,[17] is currently available to do this. MultiCoil is a two-state predictor for dimers and trimers only. Although these two oligomers account for >80% of the solved coiled-coil structures,[4] in reality, coiled coils are much more complicated: firstly, there are other higher-order and more complicated assemblies; and, secondly, within all of these, there are alternate topologies with parallel, anti-parallel and mixed arrangements of helices. SpiriCoil has advanced the state of the art in coiled-coil prediction in that it is a multistate predictor that covers all currently known possibilities. However, due to a lack of another multistate predictor, we can only test its performance comparatively against the two-state MultiCoil predictor, so we used an artificial test set constructed of only dimers and trimers, as described in Materials and Methods. The results were normalised for the unequal number of dimers/trimers in

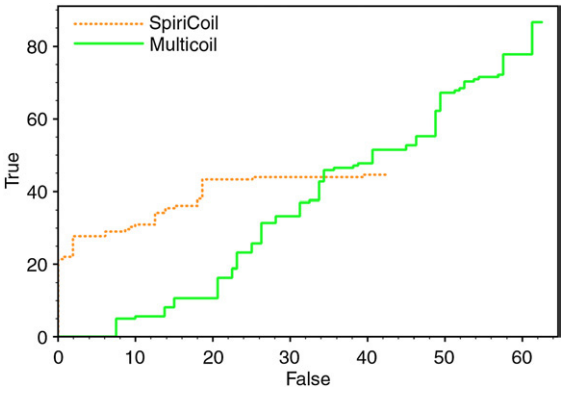

Fig. 4. Coiled-coil oligomeric state prediction benchmarking results. As in Fig. 2, the predictions made by each of the three methods are plotted from the origin in descending order of confidence. Since SpiriCoil is a multistate predictor, we count correct predictions *versus* incorrect predictions cumulatively on the *y* and *x* axes, respectively. The numbers in this figure are, however, normalised to account for the unequal numbers of dimers/trimers in the test set (counts are divided by 133/33, respectively, then multiplied by 166/2). The test set contains a total 166 possible true positives/false positives, but methods cease to report below a certain level of confidence. If MultiCoil or SpiriCoil predicts a dimer when the chain contains a dimer, or a trimer when it contains a trimer, this is a true positive. If MultiCoil or SpiriCoil predicts a trimer or a dimer when that state is not present on the chain, this is a false positive. If SpiriCoil predicts any oligomeric state other than dimer/trimer, it will be a false positive on this contrived test set; thus, where MultiCoil will get one of two true positives at random, SpiriCoil gets much fewer.

the test (otherwise, an always-dimer predictor would do significantly better than random). This test does not compare like methods, and the context must be considered when interpreting the results. For example, our training set includes only sequences with a less than 50% identity to those in

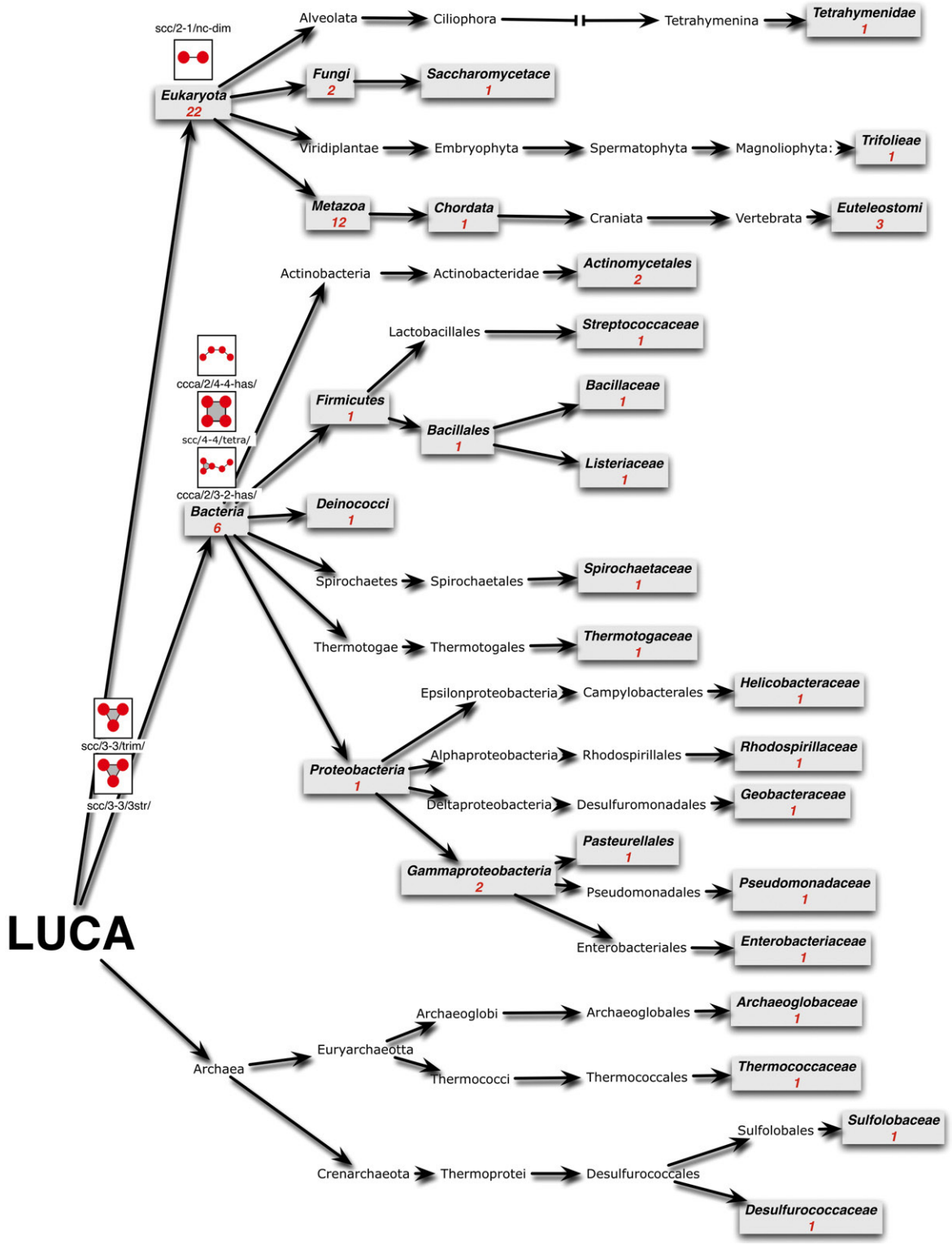

Fig. 5 (*legend on next page*)

the test set, whereas any sequence could have been used in the training of MultiCoil, which was originally trained in 1997 (we guess not subsequently updated); also, full (all state) SpiriCoil output was used. Results would appear better if it were contrived to predict only a dimer or a trimer, as the other methods do. The results in Fig. 4 show true and false assignments plotted cumulatively on the *y* and *x* axes, respectively. This shows that SpiriCoil is at least comparable with currently available oligomer state predictors on a simplified dimer/trimer test set. The method employed by SpiriCoil is fundamentally the same for all oligomeric states, and so there is no reason to believe that performance would not be the same as for dimers and trimers.

There were 54 misclassifications. Thirty-three were assigned with low confidence (*E*-values worse than 0.01) and would not be included in the genome-scale assignments that follow; the remaining 21 were neither predicted as dimer nor predicted as trimer. Of those predicted with high confidence, 9 errors were due to being assigned to families that in the training contained a vast majority of negative assignments according to SOCKET, and the remaining 12 errors were due to the sequences being in families that were not identified as coiled-coil-containing during the original annotation of the SCOP families. These families can now be added to the annotation (and will be added when SUPERFAMILY updates to version 1.75) and included in the SpiriCoil assignments without a need for retraining the models or even rescoring the genome sequences.

While coverage for MultiCoil is higher than coverage for SpiriCoil, its accuracy is lower. SpiriCoil assigned 23 of the test set to families that were known to contain coiled coils but were known to be fuzzy in their membership. In this way, the low coverage of SpiriCoil is partly due to the constraints of this test only allowing predictions to be either dimer or trimer.

Genome assignments and evolution of coiled coils

We applied the SpiriCoil procedure to all completely sequenced genomes, which totalled 1227 as of October 2009, and to other sequence sets such as UniProt (totalling 16.7 million sequences). In this way, we predicted 400,158 different coiled coils. All of the predictions and statistics on every genome are available on the SpiriCoil Web site.[1] For example, however, 2018 proteins in the human genome were predicted to contain a coiled coil (representing 4.3% of the proteome), of which 723 are more than 14 residues in length and the remainder are in the short or fuzzy class; on average, 2.53% of proteins in eukaryote genomes, 3.1% of proteins in bacteria and 1.85% of proteins in Archaea were predicted. For comparison with other genome-scale analyses for coiled coils, ARABI-COIL uses MultiCoil to identify 282 *A. thaliana* sequences that contain long coiled-coil proteins. On the other hand, SpiroCoil predicts 158 sequences, with further 124 and 803 sequences in the short and fuzzy classes, respectively, and 131 sequences in *Arabidopsis lyrata*, with further 111 and 627 short and fuzzy classes, respectively. As more coiled-coil structures are determined and deposited in the PDB, the SpiriCoil-based annotation of coiled coils in genomes will also increase.

The SpiriCoil Web site contains all of the genome assignments, a facility for users to input their own sequences for classification, and downloads of the software and all of the coiled-coil predictions in the genomes. For each coiled-coil assignment, it provides the location in the protein sequence, the oligomeric state and the most closely related structure from the PDB. It also gives a sequence alignment to that structure, as well as a 3D model made using MODELLER;[30] there is a facility to run these models back through SOCKET for analysis and comparison to the output of the experimentally determined closest structure.

Evolution of coiled-coil structure

The superfamily level in the SCOP hierarchy groups together domains that share a common evolutionary ancestor; each superfamily represents a single *de novo* evolution of a protein fold, of which there are 119 with coiled coils. Members of a superfamily are classified into families, the next level down in the hierarchy, although some superfamilies only have one possible family. The majority of the identified coiled-coil-containing superfamilies (100 of 119) are homogeneous, so all families within that superfamily contain a coiled coil. Nineteen of 119 superfamilies contain some families with coiled coils but some families without coiled coils, indicating a loss or a gain during the course of evolution

Fig. 5. An evolutionary tree of life annotated with the inception points of coiled-coil-containing superfamilies and oligomeric states. Nodes shown with the taxonomic name in rectangular boxes are superfamily inception points, with the number appearing at that point in evolution shown in red; the superfamilies corresponding to these numbers are shown in Table 4. Oligomeric state inception points are shown in square boxes above the node; the boxes contain a schematic, and the oligomeric state code is written outside the box. For example, in the top left corner, 22 superfamilies and one oligomeric state (noncanonical dimer) evolved during the formation of the 'Eukaryota' superkingdom. Note that two trimeric oligomeric states are not found in Archaea but are shown between LUCA and the two other superkingdoms (in which they are found), and the 10 remaining oligomeric states were all found in LUCA.

since the appearance of the ancestor. The vast majority of coiled coils must have been formed at the point of the *de novo* creation of the whole protein domain. There are an additional 28 superfamilies excluded from SpiriCoil despite having one or two isolated members that SOCKET identified as being coiled-coil-containing because they are vastly dominated by non-coiled-coil-containing proteins. This means that a maximum of only 47 times (19 + 28) in evolution do we observe an existing protein either lose or gain a coiled coil. We observe that coiled coils are not gained or lost by helices changing the fundamental packing of their buried residues, but they are gained or lost as decorations to the main part of the globular domain.

Coiled coils very occasionally may alter their oligomeric state in the same way, that is, not through morphological changes (e.g., trimer to tetramer, which requires existing interhelical interactions to be modified) but via addition or subtraction of helices (e.g., dimer to tetramer, which maintains existing interactions). Of the 119 true coiled-coil superfamilies, only two contain more than one different oligomeric state: In one case (the ferritins), there are many densely packed helices with two different regions that each have a fuzzy

Table 4. Coiled-coil-containing superfamilies corresponding to inception points in Fig. 5

Taxonomic group	Number of genomes	Superfamily	Oligomeric state code	Oligomeric state description
Eukaryotes				
Eukaryota	—	See Supplementary Data	Many	—
Metazoa	—	See Supplementary Data	Many	—
Chordata	26	Bcr-Abl oncoprotein oligomerization oligomerisation domain	scc/2-1/dim/anti/	Anti-parallel dimeric
Euteleostomi	38	Triple coiled-coil domain of C-type lectins	scc/3-3/trim/	Parallel trimeric
Euteleostomi	36	Chicken cartilage matrix protein	scc/3-3/trim/	Parallel trimeric
Euteleostomi	30	A tRNA synthase domain	scc/2-1/2str/anti/	Anti-parallel two-stranded
Trifolieae	1	ROP protein	scc/4-4/4str/	Four-stranded
Fungi	5	Dimerisation motif of sir4	scc/2-1/dim/para/	Parallel dimeric
Fungi	12	RNA-binding protein She2p	scc/2-1/2str/anti/	Anti-parallel dimeric
Saccharomycetace	3	VPS37 C-terminal-domain-like	scc/2-1/2str/anti/	Anti-parallel dimeric
Tetrahymenidae	1	Rotavirus nonstructural proteins	scc/2-1/dim/para/	Parallel dimeric
Eubacteria				
Bacteria	—	See Supplementary Data	Many	—
Proteobacteria	5	EspA/CesA-like	scc/2-1/dim/para/	Parallel dimeric
Gammaproteobacteria	—	See Supplementary Data	Many	—
Enterobacteriaceae	1	Positive regulator of the amidase operon AmiR	scc/2-1/dim/para/	Parallel dimeric
Pasteurellales	4	YadA C-terminal-domain-like	scc/3-3/trim/	Trimeric
Geobacteraceae	1	Fibritin	scc/3-3/trim/	Trimeric
Rhodospirillaceae	1	CO-sensing protein CooA, N-terminal domain	scc/2-1/dim/para/	Parallel dimeric
Helicobacteraceae	4	HP1531-like	scc/2-1/2str/anti/	Anti-parallel two-stranded
Thermotogaceae	7	TM0693-like	scc/2-1/dim/anti/	Anti-parallel dimeric
Spirochaetaceae	2	Variable surface antigen VlsE	scc/2-1/2str/anti/	Anti-parallel two-stranded
Deinococci	3	Seryl-tRNA synthetase (SerRS)	scc/2-1/2str/anti/	Anti-parallel two-stranded
Firmicutes	17	MW0975(SA0943)-like	scc/2-1/2str/anti/	Anti-parallel two-stranded
Bacillales	20	Hypothetical protein YfhH	scc/2-1/2str/anti/	Anti-parallel two-stranded
Listeriaceae	1	Listeriolysin regulatory protein PrfA, N-terminal domain	scc/2-1/dim/para/	Parallel dimeric
Bacillaceae	1	DinB-like	scc/2-1/2str/para/	Parallel two-stranded
Streptococcaceae	10	Spy1572-like	scc/2-1/2str/anti/	Anti-parallel two-stranded
Actinomycetales	12	Docking domain B of the erythromycin polyketide synthase (DEBS)	scc/2-1/dim/para/	Parallel dimeric
Actinomycetales	14	Rv2632c-like	scc/2-1/dim/nc-anti/	Noncanonical anti-parallel dimeric
Archaea				
Desulfurococcaceae	—	Tetrabrachion	scc/4-4/tetra/	Tetrameric
Sulfolobaceae	5	Archaeal DNA-binding protein	scc/2-1/dim/anti/	Antiparallel dimeric
Thermococcaceae	—	PF1790-like	scc/2-1/2str/para/	Parallel two-stranded
Archaeoglobaceae	1	AF0060-like	Scc/4-4/4str/	Four-stranded

The columns show the taxonomic group in which the superfamily first appeared in evolution; the number of genomes in which the superfamily is found; the name of the superfamily; the code for the oligomeric state; and the description of the oligomeric state. See Walshaw and Woolfson[2] and Moutevelis and Woolfson[4] with regard to the last two columns.

coiled coil of a different oligomeric state; sometimes SOCKET detects one, and sometimes SOCKET detects the other. In the other case of NTP pyrophosphatases, some members of the superfamily have an insertion extending two helices to form a tetramer with an existing dimer. The fuzzy set of 19 superfamilies contains more cases of varying oligomeric states, but this is because they are on the border of what SOCKET detects and is not due to major changes in structure.

Genome-level evolution of coiled coils

Genome assignments give coiled coils and their oligomeric states for all completely sequenced organisms. We used the National Center for Biotechnology Information taxonomy[31] to locate all of these organisms on a partially resolved phylogenetic species tree (Fig. 5 and Table 4). This analysis revealed the evolution of oligomeric state and individual coiled-coil-containing superfamilies. We infer the inception point of a superfamily or an oligomeric state by locating the lowest node in the tree that joins all genomes that contain it. Since a single error could cause a large change in the inception point, we carried out a manual evaluation of phylogenetic data. We checked all superfamilies and families that could be assigned to a different node in the tree, ignoring 5% of the genome assignments (i.e., allowing for a 5% error). We found that the original node assignment was unperturbed in most cases, and that they were highlighted by our evaluation merely because some parts of the tree of life are underpopulated with genomes (i.e., ignoring a single reliable prediction on the only coanoflagellate or the two dictyostelium genomes, this could in some cases move a node prediction from eukaryote to metazoa); likewise, only 7% of noneukaryote genomes are Archaea, so a minor fluctuation could tip the ratio below 5%. This suggests that the inception points have been correctly inferred. We found only seven cases where the inception point is potentially wrong: four cases where there were reliable hits to many bacterial genomes but also a single hit with a poor *E*-value to a lone eukaryote genome; three cases where there were reliable hits to a clade of bacteria but also a single hit with a poor *E*-value to an unrelated bacterium; and one case of a technical database error for a single hit (now corrected).

Remarkably, almost all oligomeric states from the periodic table of coiled-coil protein structures were present in the last universal common ancestor of life (LUCA), the only exceptions being one eukaryote-specific state, three bacteria-specific states and two states that are missing from Archaea. Similarly, the majority of superfamilies were present in LUCA, but there has been noticeable expansion in eukaryotes, especially animals. Coiled coils have arisen in evolution *de novo* 22 times between the eukaryote superkingdom being formed and the divergence into the animal, plant and fungi kingdoms. There have been 12 events between animals and chordata, 6 events during the formation of eubacteria and several other isolated events spread across all evolution, but coiled coils have only arisen *de novo* four times in archaeal lineages. This pattern of evolution is interesting but is not significantly different from that for other superfamilies with different structural characteristics. Indeed, it is consistent with recent studies of the evolution of superfamilies in general.[32]

Discussion

We present SpiriCoil (comprising a new method for coiled-coil prediction) and a resulting database of predicted coiled coils for all completely sequenced genomes. The database is available on the Web and includes annotations for the oligomeric state and 3D models for each prediction. We did not develop new technology to do this; rather, we applied a proven homology approach in a novel way, borrowing from the SUPERFAMILY resource and applying it to the coiled-coil prediction problem. Thus, the principle is to predict domains that belong to a family that we expect to contain a coiled coil. As a result, our method has one minor limitation: in predicting at the family level, we assume that all members of a family contain a coiled coil. However, it carries a major advantage that sets it apart from currently available coiled-coil prediction schemes: namely by predicting whole domains first and then locating the coiled-coil region, it makes use of valuable homology information in sequences flanking the coiled coils.

SpiriCoil outperforms existing methods in predicting the presence of coiled coils in protein sequences. Furthermore, SpiriCoil advances the state of the art in that it is the only method that can predict more complex coiled-coil oligomeric states (i.e., over and above only parallel dimers and trimers). In a somewhat contrived simple two-state (dimer/trimer) test, SpiriCoil outperforms the presently available methods. An additional and key aspect of the work presented here is the analyses of all completely sequenced genomes that it facilitates. We have found that coiled coils appear in a significant proportion of the proteins in every genome: 2.86% on average (range, 0.33–6.53%). There are currently well over 1000 genomes, and this number is growing fast; however, by integration into the SUPERFAMILY pipeline and by dynamic handling of the data, we ensure that SpiriCoil is automatically updated with new genomes as they are sequenced. SpiriCoil is currently based on SCOP version 1.73, which has now moved on to version

1.75; SpiriCoil will simultaneously update with SUPERFAMILY to version 1.75 soon, as it will for future releases.

Analysis of the coiled coils in SCOP and in the genomes has taught us that they are common elements of structure that have evolved independently many times in evolution. Coiled coils most commonly do not come about by a stepwise evolution of the packing of buried residues between helices but usually appear as part of a stable fold at its first inception. Likewise, the different oligomeric states have not evolved by morphologically changing over evolution from one to another, but rather were each independently created (often more than once). There is no perfect binary distinction between KIH packing and the lack of it; whatever criteria are chosen, there will exist proteins on the boundary of the continuum that we classify as 'fuzzy' coiled coils. This is not to say that there is no logical boundary to draw, but that this boundary will have a distribution or spread. The majority of coiled coils and almost all oligomeric states were formed before the LUCA. The few oligomeric states that have subsequently appeared did so during the formation of eukaryote and bacterial superkingdoms, as did the small number of individual coiled coils that were not present before the last universal ancestor. Since the formation of superkingdoms, only a handful of new coiled coils have evolved during the formation of the eukaryote kingdoms (concentrated in animals), and one or two in other lineages throughout evolution. Coiled coils were almost entirely forged before life existed as we see it today, almost certainly under evolutionary pressures and possibly mechanisms that are significantly different from those that are responsible for adaptation in currently living organisms. Coiled coils change little over time, and there is no evidence that they have evolved for any other reason than it is one of the possible geometric solutions to packing helices in a stable way that has a high probability of occurring by chance.

Materials and Methods

The HMM procedure for coiled-coil prediction

With the coiled coil (and oligomeric state) annotation of SCOP superfamilies and families in place, the process of predicting coiled coils using SUPERFAMILY technology was relatively straightforward, based on our previous work on SUPERFAMILY.[23,24] In brief, we constructed 190 models to extend SUPERFAMILY version 1.73 to include the 58 superfamilies in the SCOP coiled-coil class using the SAM software[33] and parameters developed for SUPERFAMILY. These models were then scored back against the sequences in SCOP version 1.73 to check for errors, which were fixed by hand.

The procedure for coiled-coil prediction in a query sequence is as follows. The query sequence is scored against the entire SUPERFAMILY library of 14,110 models representing 1831 domain superfamilies and run through the SUPERFAMILY postprocessing procedure. The result is a list of predicted domains for the query sequence including, for each, the position on the sequence, the closest structural homologue, alignment of the closest structure to the sequence, and both superfamily and family classifications. If any of the domains belongs to a superfamily or family that has been determined (above) to contain a coiled coil, then it inherits both the assignment as a coiled coil and the annotated oligomeric state. Using the alignment of the query sequence to the closest structure, coupled with the SOCKET location of the coiled coil on the structure, we map the position of the coiled coil onto the query sequence. The most similar PDB structure can be used in conjunction with this mapping to generate a 3D model using MODELLER.[30] Via the SUPERFAMILY pipeline, this procedure has been automated and rolled out to millions of sequences in over a thousand genomes. The SpiriCoil Web site provides all the assignments and a facility for users to submit their own sequences for classification online. Participation in the SUPERFAMILY pipeline means that new structures will be included via SCOP updates and newly sequenced genomes will be automatically added to SpiriCoil via SUPERFAMILY as and when they are released, without the need for any additional human action or input.

Test sets

To test as fairly as possible the performance of SpiriCoil as a coiled-coil predictor against other methods, we needed sequences that were not used to create the models. Therefore, we used SOCKET to cull a set of coiled coils from structures deposited in the PDB after SCOP version 1.73, effectively between November 2007 and March 10, 2009. These were filtered to a 95% sequence identity. This set contained 3802 protein structures, 103 of which were considered to harbor coiled coils with >14 residues according to SOCKET and using the default 7-Å cutoff for KIH interactions. At this cutoff, SOCKET can miss some coiled coils; therefore, a second set was culled with a looser cutoff of 10 Å resolution. The prediction methods tested below were said to give the following: a *true positive*, if they correctly predicted one of the 103 proteins from the first set; neither a *true positive* nor a *false positive*, if they identified one in the second (10 Å) set but not in the first set; and a *false positive*, if they identified one in neither set. This prediction test was used for the correct identification of a coiled coil in a protein, and the precise location of the coiled-coil region was not tested. The majority of coiled coils in the test set (65 of 103) belong to families that are new, since SCOP version 1.73 and the sequence similarity distribution to structures in SCOP version 1.73 can be seen in Supplementary Information (Fig. S1).

Both test sets are designed independently to best compare relative performance rather than absolute performance, and thus are not comparable to each other. To create the oligomeric state test set, we obtained parallel dimeric and parallel trimeric canonical coiled coils more than 14 amino acids in length from the CC+ database,[28] aligned them using Clustal W version 2.0[34] (maximum

gap opening and extension penalties were used to conserve the alignment of the heptad repeat) and then culled them for a 50% sequence identity (maximum) using CD-HIT.[35] The sequences were then divided according to their oligomeric state. For benchmarks, the sequence of the whole chain was submitted to the respective methods, since this best represents reality; the region of the coiled coil would not be known without the 3D structure. The set includes 133 dimers and 33 trimers (see Supplementary Data). The test sequences include 66 that are not present in SCOP version 1.73.

Implementation of other prediction methods

Currently, the best coiled-coil prediction techniques that are widely used in the field are MARCOIL[15] and Paircoil2.[13] These two techniques use different approaches to the classification technique. MARCOIL is an HMM approach where a single 64-state HMM has been trained on known coiled-coil sequences and is then used to calculate the likelihood that unseen sequences are also coiled coils. We have used all of the default MARCOIL settings and included only predictions with coils more than 14 residues in length. Paircoil2 is a technique that uses a position-specific scoring matrix in order to calculate how likely it is that a given sequence is a coiled coil. We used a window size of 28 and included only predictions with coils more than 14 residues in length. For each method, we started with the most confident prediction in the set and went through the predictions in descending order of confidence.

For oligomeric state prediction, we used MultiCoil,[17] which was originally trained in 1997. Protein sequences were submitted to the current MultiCoil Web server using default settings. For a single coiled coil, MultiCoil gives two scores at each position in the sequence (one for dimer and one for trimer); we took the ratio of the average across positions to give a single score. For SpiriCoil, we removed training sequences with a >50% sequence identity to the test sequences, which was not possible with MultiCoil, giving it a significant advantage. No prediction was made from MultiCoil in the absence of nonzero probabilities for both dimer and trimer, and no prediction was made from SpiriCoil for fuzzy and short coiled coils.

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2010.08.032

Acknowledgements

We thank Fay Moutevelis and Oli Testa for advice on CC+ and the periodic table of coiled-coil protein structures, Andrei Lupas for useful discussions, Derek Wilson for help with SUPERFAMILY, and the Biotechnology and Biological Sciences Research Council and the Engineering and Physical Sciences Research Council for funding studentships to O.J.L.R., C.T.A. and T.L.V.

References

1. Parry, D. A. D., Fraser, R. D. B. & Squire, J. M. (2008). Fifty years of coiled-coils and alpha-helical bundles: a close relationship between sequence and structure. *J. Struct. Biol.* **163**, 258–269.
2. Walshaw, J. & Woolfson, D. N. (2003). Extended knobs-into-holes packing in classical and complex coiled-coil assemblies. *J. Struct. Biol.* **144**, 349–361.
3. Lupas, A., Vandyke, M. & Stock, J. (1991). Predicting coiled coils from protein sequences. *Science*, **252**, 1162–1164.
4. Moutevelis, E. & Woolfson, D. N. (2009). A periodic table of coiled-coil protein structures. *J. Mol. Biol.* **385**, 726–732.
5. Bromley, E. H. C., Channon, K., Moutevelis, E. & Woolfson, D. N. (2008). Peptide and protein building blocks for synthetic biology: from programming biomolecules to self-organized biomolecular systems. *ACS Chem. Biol.* **3**, 38–50.
6. Lupas, A. N. & Gruber, M. (2005). The structure of alpha-helical coiled coils. *Adv. Protein Chem.* **70**, 37–78.
7. Crick, F. H. C. (1953). The packing of alpha-helices—simple coiled coils. *Acta Crystallogr.* **6**, 689–697.
8. Brown, J. H., Cohen, C. & Parry, D. A. D. (1996). Heptad breaks in alpha-helical coiled coils: stutters and stammers. *Proteins*, **26**, 134–145.
9. Hicks, M. R., Holberton, D. V., Kowalczyk, C. & Woolfson, D. N. (1997). Coiled-coil assembly by peptides with non-heptad sequence motifs. *Fold. Des.* **2**, 149–158.
10. Parry, D. A. D. (1982). Coiled-coils in alpha-helix-containing proteins—analysis of the residue types within the heptad repeat and the use of these data in the prediction of coiled-coils in other proteins. *Biosci. Rep.* **2**, 1017–1024.
11. Gruber, M., Soding, J. & Lupas, A. N. (2006). Comparative analysis of coiled-coil prediction methods. *J. Struct. Biol.* **155**, 140–145.
12. Woolfson, D. N. & Alber, T. (1995). Predicting oligomerization states of coiled coils. *Protein Sci.* **4**, 1596–1607.
13. McDonnell, A. V., Jiang, T., Keating, A. E. & Berger, B. (2006). Paircoil2: improved prediction of coiled coils from sequence. *Bioinformatics*, **22**, 356–358.
14. Berger, B., Wilson, D. B., Wolf, E., Tonchev, T., Milla, M. & Kim, P. S. (1995). Predicting coiled coils by use of pairwise residue correlations. *Proc. Natl Acad. Sci. USA*, **92**, 8259–8263.
15. Delorenzi, M. & Speed, T. (2002). An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. *Bioinformatics*, **18**, 617–625.
16. Bartoli, L., Fariselli, P., Krogh, A. & Casadio, R. (2009). CCHMM_PROF: a HMM-based coiled-coil predictor with evolutionary information. *Bioinformatics*, **25**, 2757–2763.
17. Wolf, E., Kim, P. S. & Berger, B. (1997). MultiCoil: a program for predicting two- and three-stranded coiled coils. *Protein Sci.* **6**, 1179–1189.
18. Barbara, K. E., Willis, K. A., Haley, T. M., Deminoff, S. J. & Santangelo, G. M. (2007). Coiled coil structures and transcription: an analysis of the *S. cerevisiae* coilome. *Mol. Genet.Genomics*, **278**, 135–147.
19. Newman, J. R. S., Wolf, E. & Kim, P. S. (2000). A computationally directed screen identifying interacting

coiled coils from *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA*, **97**, 13203–13208.
20. Rose, A., Manikantan, S., Schraegle, S. J., Maloy, M. A., Stahlberg, E. A. & Meier, I. (2004). Genome-wide identification of *Arabidopsis* coiled-coil proteins and establishment of the ARABI-COIL database. *Plant Physiol.* **134**, 927–939.
21. Rose, A., Schraegle, S. J., Stahlberg, E. A. & Meier, I. (2005). Coiled-coil protein composition of 22 proteomes—differences and common themes in subcellular infrastructure and traffic control. *BMC Evol. Biol.* **5**, 66.
22. Liu, J. F. & Rost, B. (2001). Comparing function and structure between entire proteomes. *Protein Sci.* **10**, 1970–1979.
23. Gough, J. & Chothia, C. (2002). SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res.* **30**, 268–272.
24. Gough, J., Karplus, K., Hughey, R. & Chothia, C. (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.
25. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP—a Structural Classification of Proteins Database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540.
26. Andreeva, A., Howorth, D., Chandonia, J. M., Brenner, S. E., Hubbard, T. J. P., Chothia, C. & Murzin, A. G. (2008). Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.* **36**, D419–D425.
27. Walshaw, J. & Woolfson, D. N. (2001). SOCKET: a program for identifying and analysing coiled-coil motifs within protein structures. *J. Mol. Biol.* **307**, 1427–1450.
28. Testa, O. D., Moutevelis, E. & Woolfson, D. N. (2009). CC plus: a relational database of coiled-coil structures. *Nucleic Acids Res.* **37**, D315–D322.
29. Wilson, D., Madera, M., Vogel, C., Chothia, C. & Gough, J. (2007). The SUPERFAMILY database in 2007: families and functions. *Nucleic Acids Res.* **35**, D308–D313.
30. Fiser, A. S. & Sali, A. (2003). MODELLER: generation and refinement of homology-based protein structure models. *Macromol Crystallogr. D*, **374**, 461–491.
31. Sayers, E. W., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V. *et al.* (2009). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **37**, D5–D15.
32. Chothia, C. & Gough, J. (2009). Genomic and structural aspects of protein evolution. *Biochem. J.* **419**, 15–28.
33. Karplus, K., Barrett, C. & Hughey, R. (1998). Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, **14**, 846–856.
34. Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H. *et al.* (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.
35. Li, W. Z. & Godzik, A. (2006). CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.