Published in final edited form as:
*Bioessays*. 2012 December ; 34(12): 1060–1069. doi:10.1002/bies.201200116.

# When a domain isn't a domain, and why it's important to properly filter proteins in databases:

## Conflicting definitions and fold classification systems for structural domains makes filtering of such databases imperative

**Clare-Louise Towse** and **Valerie Daggett**
Department of Bioengineering, University of Washington, Seattle, WA 98195-5013, United, States

## Summary

Membership in a protein domain database does not a domain make; a feature we realized when generating a consensus view of protein fold space with our Consensus Domain Dictionary (CDD). This dictionary was used to select representative structures for characterization of the protein dynameome: the Dynameomics initiative. Through this endeavor we rejected a surprising 40% of the 1695 folds in the CDD as being non-autonomous folding units. Although some of this was due to the challenges of grouping similar fold topologies, the dissonance between the cataloguing and structural qualification of protein domains remains surprising. Another potential factor is previously overlooked intrinsic disorder; predicted estimates suggest 40% of proteins to have either local or global disorder. One thing is clear, filtering a structural database and ensuring a consistent definition for protein domains is crucial, and caution is prescribed when generalizations of globular domains are drawn from unfiltered protein domain datasets.

**Keywords**

computational biology; protein structure; dynameome; intrinsic disorder

## Introduction

The interpretation of "protein domain" is highly dependent upon context, yet in all definitions a domain is a region of an amino acid sequence with features that are repeated throughout the protein kingdom. These features can be a region of conserved primary structure [1], part of a sequence known to confer function, or defined by the boundaries within which the sequence can form a structural unit. A structural domain is loosely defined as a unit that can exist independently when excised from the full protein or complex. More specifically, a domain is an independently structured unit capable of autonomous folding. In many cases function is related to structure, and the boundaries of a functional domain can correlate with those of the structural domain, with sequence conservation in both the structural and functional sense. However, in the structural context, protein domain families need not have high sequence similarity; they require only similarity in the tertiary structure of that "fold". Grouping structures based on similarity, or performing a geometric alignment of conserved structure, is decidedly more challenging than the alignment of a protein sequence. Two proteins with a similar arrangement of secondary structure may not have the same chain directionality through the secondary structure elements, and they may contain non-conserved regions that are significantly different. Figure 1A illustrates this scenario for five domains from the ferredoxin-like fold family. The commonality in the arrangement of two α-helices packed against a β-sheet is apparent, yet there is extraneous structure and variation in the lengths and alignment of loop regions. Hence, defining a domain can be a complex matter and is susceptible to much uncertainty and inconsistency.

The origin of classifying structural units into domains came following the solution of very similar structures starting with myoglobin in 1958 [2] and haemoglobin in 1960 [3]. It was clear even from these early low-resolution structures that there were common structural elements. The simplest of these show collections of secondary structure found together in repeated arrangements [4] that appear in all members of a given fold family. Domains themselves, can be classified based on the overall structure observed, all-α, all-β or a mixture of α/β or α+β. Examples within these broad classes can be found in Figure 1B: all-β, immunoglobulin-like;α+β, flavodoxin-like; α/β, TIM barrel; and all-α, three-helical bundle. Today, identification of a novel domain structure, or fold, is rare [5]; since 2008 the RCSB Protein Data Bank (PDB, www.pdb.org) has not registered the deposition of any new folds [6]. Despite the huge number of folds theoretically possible [7], many believe that protein fold space is nearly complete, at least for single-domain globular proteins [8–11].

Although similarities between protein structures are often visibly apparent, classifying protein structure suffers from three main hurdles: deciding where the boundaries of a domain lie, how to group similar structures together, and when a structure is ‘too’ different to be part of a group [11]. The first comprehensive attempt to group protein domains by structural similarity was in 1976 when patterns were categorized across a set of 31 globular proteins [4]. Since then, there have been three leading databases that have specialized in categorization of protein structure into fold families: SCOP [12], CATH [13] and Dali [14].

The Structural Classification of Proteins database (SCOP) started as a manual effort, with visual inspection used to identify domains and classify them based on evolutionary relationships [12]. The structures were first placed into all-α, all-β or mixed αβ classes based on the overall secondary structure content, then grouped by shared function or structural features irrespective of their sequence similarity. Next, they were grouped based on sequence similarity, and finally by the nature of the conserved topologies. Due to increasing speed with which protein structures were being solved, this eventually became a partially automated effort along with refinement of the classification definitions [15]. CATH derives its name from the classification system it uses: C, class; A, architecture; T, topology; H, homologous superfamily [13]. The systems uses fully automated sequence alignments, structure comparisons and domain definitions, with homology detected at the 35% sequence identity limit. Dali first determines the presence of a domain depending on how compact it is and uses a neural network to perform fully automated domain classifications [14].

Providing further testament to the challenge of categorizing structures, there are inconsistencies regarding how many of the domain structures have been siphoned into domain families by these protein domain databases. To perform a large-scale assessment across protein fold space, such conflict in the assignment of a structure to one domain family or another presents a problem. For this reason, the definition and classification of protein structures became a key part of the Dynameomics project: an initiative to simulate and catalogue the dynamics of representatives of all known protein folds (Fig. 1B) [16, 17].

Given the structural similarity observed through protein fold space, Dynameomics focuses on representatives of protein folds rather than all known protein structures, making the problem more tractable while still achieving sampling of fold space. In contrast, other ‘big’ science projects, such as the Human Genome and Human Proteome projects require waiting for the entire human genome and proteome to be catalogued [18]. Dynameomics began over a decade ago and 807 proteins of different topologies, representing essentially all known globular protein folds, have now been simulated at both room and high temperatures, totaling hundreds of terabytes of data (Fig. 1B). At the beginning of this effort, it was clear that discrepancies between the different protein domain databases presented a problem and some uniformity needed to be established. To catalogue the dynamics of all known protein domains we needed a consensus view of fold space.

## The Consensus Domain Dictionary consolidated the visible landscape of protein fold space

The motivation behind generation of a consensus domain dictionary (CDD) [19, 20] was to gather a set of representative protein structures that could be used to systematically investigate all of protein fold space and determine the principles of protein dynamics and folding. This structure-based dictionary should not be confused with the Conserved Domain Database [1, 21] that categorizes the primary sequences of protein domains from an evolutionary standpoint.

The collation of all identified structural domains currently in SCOP, CATH and Dali into a consensus set was initially done in 2003 using a metadata approach [19]. This was then updated in 2009 to include any new protein folds discovered in that interim period [20]. The total number of metafolds increased by 595, reflecting not just newly discovered folds but also the refinement of structural classifications that occurred during this six year period [15, 22]. Consequently, there were some domains made obsolete, re-delineation of some domain boundaries, as well as a merging of domains and metafolds within the v2009 CDD. Once complete, there had been inclusion of 976 “new” metafolds, composed entirely of “new” domains that were not present in the v2003 CDD.

After assimilating and filtering the information contained within our v2009 CDD, the final Dynameomics dataset contained 807 representative structures taken from metafolds that encompass 97% of all currently known protein structures (Fig. 2) [16, 17, 23].

## Numerous classified domains were not “traditional” domains

As the goal of the Dynameomics initiative has been to perform atomistic molecular dynamics (MD) simulations of the native state and unfolding pathways of representative structures for all known globular protein domains, some of the metafolds included in the CDD were superfluous to our needs. Any non-globular proteins (*i.e.* transmembrane or fibrous proteins) were not included, but these were few (Fig. 2). Due to the limits of our modeling methods and computational power we restricted the study to the part of the proteome where domains were less than 450 residues and rejected those with multiple or large cofactors for which we had no parameters. These restrictions were seconded by the facts that few domains are larger than 450 residues [24] and, although parameters could be developed for the multiple or large cofactors, many domains with such cofactors are really a polypeptide wrapped around a cofactor, such that there is little conventional structure. Other reasons for rejection were experimental structures of suboptimal quality or with large regions of missing coordinates (Fig. 2). Further, there were 19 NMR structures that were initially included but when modeled were found to have unstable native states. All but five, which had alternative X-ray crystal structures that were substituted and simulated successfully, were removed from the study. The starting structures for these 14 rejected simulations were designated questionable. In total, 888 protein domains were rejected in response to the requirements of the Dynameomics project (Fig. 2). Of these 888 rejected metafolds, 672 were not simulated because they are not structural protein domains. These were the metafolds we rejected because their structures were irregular with little secondary structure, structural units that were unlikely to fold autonomously, or composite domains, where the domain of interest was interrupted by a second structural domain, *i.e.* discontinuous (Fig. 2). The v2009 CDD definitions and representative structures of the 1695 metafolds, along with the 807 representatives qualified as “suitable” domains and successfully simulated, can be accessed at http://www.dynameomics.org/external/Targetlist/index.aspx. Overall, the majority of the rejected metafolds were eliminated because they are unable to be designated as autonomous domains or folds.

Here we revisit the 888 rejected metafolds, as there are subsets that warrant further inspection, especially in light of the expanding knowledge of the existence of proteins with local or global regions that lack structure, the so-called intrinsically disordered proteins (IDPs) [25]. These subsets may assist us in adapting our definition of domains to include what now appears to be a continuum of structures from disordered to highly regular structure. The structures we rejected because they had regions of atoms that were experimentally invisible, were non-autonomous folders, or had irregular and unstable structure, could be part of the unfoldome [26]. To avoid including further complexity at this point, we continue to ignore transmembrane and fibrous proteins and focus on globular protein or constituent domains less than 450 residues in length. We survey here the resulting dataset of 755 rejected metafolds (Fig. 2) by reexamining discarded simulations and performing disorder predictions using the DISOPRED2 classifier [27].

In the latest rendition of the CDD, it was noted as surprising that 40% of the metafolds in the CDD were not autonomous structural units and were therefore not believed to be “real” domains [20]. However, in light of the recent progress made in understanding disordered proteins, which has led to the prediction that 40 – 50% of mammalian proteins contain disordered regions greater than 30 residues in length [26, 28], maybe this isn’t so surprising after all. The PDB itself has been surveyed for intrinsic disorder, with one conclusion being that completely ordered proteins are not that abundant in this structural database [29]. Approximately 40% of the structures in the PDB have short disordered regions, defined as a continuous region between 10 and 30 residues long [29]. It is not so unexpected then that, based on predictions made here using DISOPRED2 [27], some 34% of the 755 rejected metafolds contain putative disordered regions greater than 10 residues in length. Although undoubtedly some of these targets were quite rightly rejected on the basis of not being viable structural domains, it does appear that some may fall into the IDP category. As DISOPRED2 predictions are more biased towards identifying short disordered regions within globular proteins, this 34% is potentially an underestimate [27, 30].

## Instability could be attributable to disorder

Instability during an MD simulation is generally ascribed to structural or methodological errors, but this implies that domains are expected to always be as structurally cohesive as the model calculated from the experimental data. Initially, we had selected representatives from 821 metafolds as our target list. For a simulation to be considered stable it had to pass our quality control measures [16, 17]. However, 19 of these domains were found to be unstable, with some unpacking of the native structure and loss of secondary structure. All 19 starting structures were determined using NMR; five had alternative X-ray crystal structures and simulations beginning with these substitute structures were stable. Although none of these unstable targets has yet to be entered in the DisProt database of disordered structures [31], ten of these metafolds have putative or confirmed regions of disorder greater than ten residues. For half of these, the disorder forms a substantial part of the domain (>40%) and in two of the larger domains, longer disordered regions are observed (PDB ID: *1wgo* and *7hsc*) even though these account for a small fraction of the domain. Thus, while instability in an MD simulation can be due to methodological problems or a problematic starting structure, it can also result from true dynamically disordered regions.

## Unstable regions had disordered sequence signatures

Six of the targets are predicted to have disordered regions that overlap with unstable regions (Fig. 3): PDB IDs: *1vpu, 1t23, 1q3j, 1fu9, 1k0h, 7hsc*. The first, the HIV-1 protein U (Vpu) is an auxiliary protein that has already been noted within the intrinsically disorder community [32] as potentially one of the disordered proteins within the HIV-1 proteome. Vpu is a transmembrane protein, but the cytoplasmic domain (PDB ID: *1vpu*) is a globular domain [33]. This domain has a high content of acidic residues and an ampipathic helix adjacent to the transmembrane region where disorder was predicted, both here and over a more extended region by others (Fig. 3A) [32]. These predictions correlate with the conformational variation and loss of structure exhibited in the simulation; this is evident both in the MD ensemble of conformations and the final resulting structure (Fig. 3A). Another example is the representative for the chromosomal protein MC1 fold (PDB ID: *1t23*). This domain has regions that lose structure and unpack from the surface of the remaining folded β-sheet; it is these regions that are predicted to be disordered (Fig. 3). The only discrepancy between the predicted disordered region and the simulation is the helix, which remained structured, although it did unpack from the surface of the β-sheet.

## Examples of genuine instability

There are a number of simulations that exhibited instability in regions predicted to be disorderd; but the agreement is likely coincidental as the simulations remain questionable because of the choice of starting structure and presence of cofactors and disulfide bonds. For example, simulations of the heat shock chaperone protein Hsc70 used the deposited average structure generated from an NMR ensemble (PDB ID: *7hsc*) and show the front-facing β-strands to be particularly unstable (Fig. 3B). Only one structure was deposited and it is a highly energy minimized average structure. This simulation was an oversight given the likelihood for artificiality in the average structure that, while reflecting the NMR ensemble well, predisposes it to instability as an isolated conformation. Although DISOPRED2 predicts a disordered region longer than 20 residues that coincides with the hinge in the helical region (Fig. 3B) where there are large-scale movements upon binding [34], it is difficult to judge whether it is the starting structure itself or the potential disorder propensity that is responsible for the instability we observed. Other questionable cases are the CCHC-type zinc finger domain (PDB ID: *1fu9*) and the knottin-like antifungal peptide Alo-3 (PDB ID: *1q3j*). In the absence of the metal ion, CCHC-type zinc fingers have been reported to be disordered [35], and they have primary sequence content biases typical of IDPs [36]. Hence, the predicted and MD observed disorder at the termini of *1fu9* is expected. However, the simulation of *1fu9* included a bound zinc ion, which induces structure [37], and, although having the characteristics of a disordered protein, this domain should have remained stable. The knottin fold of Alo-3 has a long flexible loop between strands β1 and β2 and, like many antimicrobial peptides, has an overall cationic charge. Again, the predicted disorder coincidental with the flexible loop was expected, given that other antimicrobial peptides have unstructured or extended structures under some conditions [38]. In this case, the secondary structure was lost rapidly during the simulation but the fold was retained due to the disulfide bonds.

## Structure or the lack of structure can be context-dependent

Two proteins from bacteriophage λ, protein W (PDB ID: *1hyw*) and FII protein (PDB ID: *1k0h*), with confirmed unstructured regions were also amongst those we noted to have disordered regions [39, 40]. In both cases we did not simulate the unstructured regions at the termini. However, in the case of protein FII, there was a second unstructured region detected experimentally that we did simulate. This second region was also predicted to be disordered and is located in the central loop between two β-strands (Fig. 3B).

The other unstable targets with shorter contiguous predicted disordered regions were found to have experimentally observed unstructured regions associated with them, often on the edges of the domain, for example the PKD domain of the VPS10 receptor, (PDB ID: *1wgo*). These domains had either regions with random coil chemical shifts or missing coordinates in the PDB files for sets of terminal residues [39, 41]. Hence, in most cases the domains were simulated without these segments. It is plausible that some interactions between the fold and the unstructured regions are required to maintain the stability in these instances [17]; this could be the reason behind the extreme loss of secondary structure exhibited by *1k0h* that was simulated with the unstructured N-terminus removed (Fig. 3B).

Although poorly refined structures, potentially artifactual average NMR structures, or over-restrained structures could be a contributing factor to the instability observed, there was some correlation with the location of the putative disordered regions. Encouragingly, some of the targets with unstable simulations have since become established IDP cases.

## Irregular structures are predominantly motifs

The sequence lengths of the "irregular" domains ranged from 31 to 412 residues with most clustered between 30 and 40 residues. Hence, many of the 87 "irregular" domains (Fig. 2) are small peptides and appear to consist of supersecondary structure (Fig. 1A) [4]. Some show truncated βαβ motifs, othersβββ and α folding units (Fig. 4A). It is difficult to know how to assign these structures when they represent the structural motifs found in most all domain structures.

87% of the "irregular" folds were solution NMR structures with significant regions lacking in well-defined secondary structure. Approximately half of these folds have 30% or greater of their sequences predicted to be disordered, but with many being relatively short sequences. There are few that have predicted disordered regions exceeding 20 contiguous residues. Four of the irregular targets are currently listed in the DisProt database (Fig. 4A): the CCHCC (PDB ID: *1pxe*), connexin43 (PDB ID: *1r5s*), knottin (PDB ID: *1ha9*), and cysteine α-hairpin motif (PDB ID: *1u97*) folds. Only one fold, aptly named partially disordered protein At2g23090, has over 30 contiguous residues predicted as disordered, greater than 50% of its sequence, which corresponded with a lack of experimentally determined secondary structure (PDB ID: *1wvk*) (Fig. 4A). However, although the coordinates were deposited in the PDB in 2004, there is no reference to an associated publication, nor is the protein included in DisProt. It is, however, a member of the SERF family (InterPro: IPR007513) [42]; a family of proteins that contain a high content of charged residues (aspartic acid, glutamic acid, lysine and arginine) that could be an interesting collection of potential IDPs.

## Missing atomic coordinates tend to reflect disorder

Experimental determination of protein structures is not without its flaws. One issue that is prevalent throughout the PDB for X-ray crystal structures is missing atomic coordinates where electron density is not detected or highly diffuse [27]. Most of the regions lacking electron density in crystal structures have amino acid compositions predicted to be disordered [29], and it is possible that a fraction of the observed secondary structure may have been induced by crystallization in complex with binding partners. This group of 85 "rejects" consisted entirely of such X-ray crystal structures, with missing coordinates for regions of seven residues or longer. The majority of these folds are single domains between 100 and 200 residues. Hence, a significant number of these rejected targets are expected to contain local disorder.

In total, 46% of these 85 folds contained some disorder. Eight targets had at least 30 contiguous residues predicted to be disordered (Fig. 4B), with another 31 with shorter, predicted disordered regions, the majority of which correlated with the missing coordinates. Cases of particular interest were those predicted to have greater than 30 contiguous disordered residues and that are present in DisProt; there were three such targets in DisProt (PDB ID: *1bo1*, *1qwy*, *1y8q*).

For all eight targets with significant predicted disorder, the disorder coincided with the missing atomic coordinates. In a number of cases, regions outside the missing coordinates were also predicted to be disordered and correlated with lower secondary structure content (Fig. 4B). In many cases the disorder was mentioned in the corresponding literature. For example, the disorder propensity of a domain of the 17kDa protein (Skp) with an outer membrane protein H (OmpH)-like fold (PDB ID: *1u2m*) has been investigated previously and is in agreement with the prediction obtained here [43]. Disordered residues were reported in the unbound form of the C-terminal domain of the pepsin inhibitor-3 (PI-3) protein (PDB ID: *1f32*) and some were noted to become ordered upon formation of a complex with pepsin, although some disorder still remained in the complex [44].

Also interesting are the occurrences of missing coordinates that are not predicted to be disordered and not registered in DisProt. An example is the protein kinase-like fold of phosphatidylinositol phosphate kinase type IIβ (PDB ID: *1bo1*). There are two areas of the structure missing coordinates (Fig. 4B). DISOPRED2 predicts only one of the two missing segments to be disordered. The lag in inclusion into DisProt is likely a consequence of the manual curation required [45]. As for the disagreement with the predictions, it has been noted that DISOPRED2 often predicts some regions with missing coordinates to be ordered and, in some cases, they have been revealed to be regions with transient structure [30].

Although having missing electron density is the very reason many of these targets have been reported to contain disordered regions, it appears likely that the majority of these targets are typical globular domains with regions of local disorder.

## Targets rejected for being non-autonomous domains harbor the most intrinsic disorder

Non-autonomous domains, by their very nature of being unable to assume structured folds independently, are a logical group to scrutinize for disorder. The 585 folds labeled as “non-autonomous”, are not significantly larger than those folds in the other groupings, with only 11 containing over 450 residues. However, they typically form part of large complexes (PDB ID: *1eq2* in Fig. 4C). Most of these non-autonomous folds, being structured in the context of a huge complex or in the presence of binding partners, were rejected because they had large regions buried within a complex making them unlikely to form stable structure in isolation. Another 86 folds are composite domains where the structural unit is interrupted by structural elements from a second domain or chain. Examples of these folds are the Rossman, immunoglobulin-like, and citrate synthase domain 1 folds: PDB IDs *1eq2*, *1i31* and *1css*, respectively, in Figure 4C. Although it may seem strange to have structural units with spatially separated sequences, remember that these composite domains are present in the underlying databases as a consequence of the different methods used for categorization. For example, at one of the top-levels in the classification hierarchy of CATH, structures are grouped where they have a similar spatial arrangement of secondary structure elements, irrespective of chain directionality or continuity [46].

Disorder predictions were performed on the 488 continuous folds; the fragmented nature of the 86 composite domains prohibited meaningful sequence analysis. Some 182 of the 488 targets had predicted disordered regions at least ten residues in length and approximately 70 had missing coordinates, some coincidental with disordered regions as observed above for the crystal structures. Greater disorder was indicated in 43 of these folds, which were either listed in DisProt or had predictions for more than 30 contiguous disordered residues, or both. Three of these in particular, the C-terminal domain of topoisomerase II (*1bjt*), cAMP-responsive binding protein (CBP, *1kbh*) domain, and heat shock locus HSLU (*1ht2*), have disorder predictions that correlate with regions already highlighted in DisProt (Fig. 4C).

The CBP domain (PDB ID: *1kbh*) is an established IDP [47]. A first of its kind, this fold representative was one of two domains that are disordered in isolation but become structured upon forming a heterodimer. The other domain is the receptor activator (ACTR) domain, which exists as a molten globule until it forms a complex with the CBP domain [48]. Although shown here in the complexed, structured form with the ACTR domain shaded in gray (Fig. 4C), NMR indicates that the unbound form of the CBP domain is 100% disordered [48]. The C-terminal domain of topoisomerase II (PDB ID: *1bjt*) has a region of missing coordinates representing an acidic region sensitive to protease degradation, leading to the suggestion that it is disordered [49, 50]. The heat shock locus HSLU domain (PDB ID: *1ht2*) has two predicted disordered regions, one that matches with missing coordinates in the X-ray crystal structures [51, 52] and both are listed in DisProt.

Coupled folding and binding is synonymous with the presence of disorder in unbound states. Accordingly, these rejected non-autonomous folds have a high incidence of known and putative disordered regions.

## Recasting the Definition of a Domain

Much of what was believed for the last 50 years regarding the relationship between structure and function is now in question as more and more proteins with a local or global lack of structure are identified [25]. The dogma that the native, biologically active state is folded and structured is biased because of the limitations of experimental techniques. Although there were some exceptions to this rule, they were believed to be just that. However, there has been increasingly more evidence accruing over the last years that all is not what it first seemed [25].

It is possible that as we learn more, the folds we previously rejected for not being *bona fide* domains may be considered as such and we should be open to adapting or expanding the definition of domains. Many disordered proteins have some fold or structure, albeit less compact than the “typical” native states of globular proteins, with many able to form different conformations with different binding partners. Perhaps, this is an indication we should start to consider “dynamic domain” families? After all, conservation of disordered regions has been demonstrated [53]. Moreover, some disordered regions have already shown attributes consistent with domains, leading to consideration of disordered domain classification [54]. However, at the start of such a revolution in understanding protein structure, or non-structure, it becomes even more important to be able to confidently define domain boundaries and group similar folds together before including ‘dynamic’ or ‘disordered’ protein fold families. If the complexity of attempting to categorize well-structured proteins into domain families has taught us anything, we can be sure that as the gaps in protein fold space are filled, defining a domain will only become more challenging.

## Conclusions

In our Dynameomics project we generated a consensus set of protein folds and selected targets based on the physical attributes of what constituted a globular structural domain. In the course of this work, it became clear that – in some cases - there was much ambiguity as to what qualifies as a domain and what does not. Many of the structures assigned to domain families in CATH, SCOP and Dali were not strictly domains in a structural sense. When we updated our CDD in 2009, the field of intrinsically disordered proteins was still an emerging field. Inspection of our rejected targets here revealed that many contain disordered regions. Fortunately, our consensus approach, along with manual inspection, eliminated many of the domains with confirmed or putative disorder. This reiterates how important it was to survey and re-qualify our consensus data prior to Dynameomics.

Re-examination of our rejected targets here not only demonstrates the importance of filtering structural databases and applying a consistent descriptor for structural domains, it also raises questions for the near future. At some point we will gather sufficient knowledge to consider categorization of IDPs into domain families, a somewhat daunting notion. The challenge of structural domain categorization is only going to become more complex. The proteome is now seen as existing on a continuum of varying structural content, with structure not necessarily present under physiological conditions, and disorder and order interchangeable states dependent upon the environment and protein interactions.

So when is a domain really a domain? And if the native state isn’t always the folded state, which state do we use to determine membership in a domain family? Although our definition of a domain may need to be adapted and expanded, we can take comfort that, for now, when focusing on typical globular domains one only needs to ensure appropriate filtering of the data.

## Acknowledgments

We thank Dr. Andrew Moore for suggesting that we delve again into this fascinating facet of structural biology, especially in light of the changing paradigms driven by increasing evidence of the distinct roles that order and disorder play. Thanks also go to R. Dustin Schaeffer who guided Clare-Louise in understanding his previous work in this area. We are grateful for support from the NIH (GM50789). Protein images were rendered using the UCSF Chimera package [55].

## References

1. Marchler-Bauer A, Anderson JB, Derbyshire MK, DeWeese-Scott C, et al. CDD: a conserved domain database for interactive domain family analysis. Nucleic Acids Res. 2007; 35:D237–40. [PubMed: 17135202]

2. Kendrew J, Bodo G, Dintzis H, Parrish R, et al. A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. Nature. 1958; 181:662–6. [PubMed: 13517261]

3. Perutz MF, Rossmann MG, Cullis AF, Muirhead H, et al. Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-Å resolution, obtained by X-ray analysis. Nature. 1960; 185:416–22. [PubMed: 18990801]

4. Levitt M, Chothia C. Structural patterns in globular proteins. Nature. 1976; 261:552–8. [PubMed: 934293]

5. Grant A, Lee D, Orengo C. Progress towards mapping the universe of protein folds. Genome Biol. 2004; 5:107. [PubMed: 15128436]

6. Berman HM, Westbrook J, Feng Z, Gilliland G, et al. The protein data bank. Nucleic Acids Res. 2000; 28:235–42. [PubMed: 10592235]

7. Taylor WR, Chelliah V, Hollup SM, MacDonald JT, et al. Probing the “dark matter” of protein fold space. Structure. 2009; 17:1244–52. [PubMed: 19748345]

8. Skolnick J, Zhou H, Brylinski M. Further evidence for the likely completeness of the library of solved single domain protein structures. J Phys Chem B. 2012; 116:6654–64. [PubMed: 22272723]

9. Kihara D, Skolnick J. The PDB is a covering set of small protein structures. J Mol Biol. 2003; 334:793–802. [PubMed: 14636603]

10. Skolnick J, Arakaki AK, Lee SY, Brylinski M. The continuity of protein structure space is an intrinsic property of proteins. Proc Natl Acad Sci USA. 2009; 106:15690–5. [PubMed: 19805219]

11. Schaeffer RD, Daggett V. Protein folds and protein folding. Protein Eng Des Sel. 2011; 24:11–9. [PubMed: 21051320]

12. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol. 1995; 247:536–40. [PubMed: 7723011]

13. Orengo CA, Michie AD, Jones S, Jones DT, et al. CATH a hierarchic classification of protein domain structures. Structure. 1997; 5:1093–109. [PubMed: 9309224]

14. Dietmann S. A fully automatic evolutionary classification of protein folds: Dali Domain Dictionary version 3. Nucleic Acids Res. 2001; 29:55–7. [PubMed: 11125048]

15. Andreeva A, Howorth D, Brenner SE, Hubbard TJP, et al. SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res. 2004; 32:D226–9. [PubMed: 14681400]

16. Van Der Kamp MW, Schaeffer RD, Jonsson AL, Scouras AD, et al. Dynameomics: A comprehensive database of protein dynamics. Structure. 2010; 18:423–35. [PubMed: 20399180]

17. Beck DAC, Jonsson AL, Schaeffer RD, Scott KA, et al. Dynameomics: mass annotation of protein dynamics and unfolding in water by high-throughput atomistic molecular dynamics simulations. Protein Eng Des Sel. 2008; 21:353–68. [PubMed: 18411224]

18. Lander ES, Linton LM, Birren B, Nusbaum C, et al. Initial sequencing and analysis of the human genome. Nature. 2001; 409:860–921. [PubMed: 11237011]

19. Day R, Beck DAC, Armen RS, Daggett V. A consensus view of fold space: Combining SCOP, CATH, and the Dali Domain Dictionary. Protein Sci. 2003; 12:2150–60. [PubMed: 14500873]

20. Schaeffer RD, Jonsson AL, Simms AM, Daggett V. Generation of a consensus protein domain dictionary. Bioinformatics. 2011; 27:46–54. [PubMed: 21068000]

21. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, et al. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res. 2010; 39:D225–9. [PubMed: 21109532]

22. Andreeva A, Howorth D, Chandonia J-M, Brenner SE, et al. Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res. 2008; 36:D419–25. [PubMed: 18000004]

23. Benson NC, Daggett V. Dynameomics: Large-scale assessment of native protein flexibility. Protein Sci. 2008; 17:2038–50. [PubMed: 18796694]

24. Islam SA, Luo J, Sternberg MJE. Identification and analysis of domains in proteins. Protein Eng Des Sel. 1995; 8:513–26.

25. Uversky VN. Intrinsically disordered proteins from A to Z. Int J Biochem Cell Biol. 2011; 43:1090–103. [PubMed: 21501695]

26. Uversky VN. The mysterious unfoldome: structureless, underappreciated, yet vital part of any given proteome. J Biomed Biotech. 2010; 2010:568068.

27. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, et al. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol. 2004; 337:635–45. [PubMed: 15019783]

28. Fink AL. Natively unfolded proteins. Curr Opin Struc Biol. 2005; 15:35–41.

29. Le Gall T, Romero PR, Cortese MS, Uversky VN, et al. Intrinsic disorder in the Protein Data Bank. J Biomol Struct Dyn. 2007; 24:325–42. [PubMed: 17206849]

30. Dosztányi Z, Csizmók V, Tompa P, Simon I. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J Mol Biol. 2005; 347:827–39. [PubMed: 15769473]

31. Vucetic S, Obradovic Z, Vacic V, Radivojac P, et al. DisProt: a database of protein disorder. Bioinformatics. 2005; 21:137–40. [PubMed: 15310560]

32. Xue B, Mizianty MJ, Kurgan L, Uversky VN. Protein intrinsic disorder as a flexible armor and a weapon of HIV-1. Cell Mol Life Sci. 2012; 69:1211–59. [PubMed: 22033837]

33. Willbold D, Hoffmann S, Rösch P. Secondary structure and tertiary fold of the human immunodeficiency virus protein U (Vpu) cytoplasmic domain in solution. Eur J Biochem. 1997; 245:581–8. [PubMed: 9182993]

34. Morshauser RC, Hu W, Wang H, Pang Y, et al. High-resolution solution structure of the 18 kDa substrate-binding domain of the mammalian chaperone protein Hsc70. J Mol Biol. 1999; 289:1387–403. [PubMed: 10373374]

35. Berkovits HJ, Berg JM. Metal and DNA binding properties of a two-domain fragment of neural zinc finger factor 1, a CCHC-type zinc binding protein. Biochemistry. 1999; 38:16826–30. [PubMed: 10606515]

36. Uversky VN, Gillespie JR, Fink AL. Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins. 2000; 41:415–27. [PubMed: 11025552]

37. Matthews JM, Kowalski K, Liew CK, Sharpe BK, et al. A class of zinc fingers involved in protein-protein interactions. Eur J Biochem. 2001; 267:1030–8. [PubMed: 10672011]

38. Falla TJ, Karunaratne DN, Hancock RE. Mode of action of the antimicrobial peptide indolicidin. J Biol Chem. 1996; 271:19298–303. [PubMed: 8702613]

39. Maxwell KL, Yee AA, Booth V, Arrowsmith CH, et al. The solution structure of bacteriophage λ protein W, a small morphogenetic protein possessing a novel fold. J Mol Biol. 2001; 308:9–14. [PubMed: 11302702]

40. Maxwell KL, Yee AA, Arrowsmith CH, Gold M, et al. The solution structure of the bacteriophage λ head tail joining protein, gpFII. J Mol Biol. 2002; 318:1395–404. [PubMed: 12083526]

41. Fukushima K, Kikuchi J, Koshiba S, Kigawa T, et al. Solution structure of the DFF-C domain of DFF45/ICAD. A structural basis for the regulation of apoptotic DNA fragmentation. J Mol Biol. 2002; 321:317–27. [PubMed: 12144788]

42. Hunter S, Jones P, Mitchell A, Apweiler R, et al. InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res. 2012; 40:D306–12. [PubMed: 22096229]

43. Paliy O, Gargac SM, Cheng Y, Uversky VN, et al. Protein disorder is positively correlated with gene expression in Escherichia coli. J Proteome Res. 2008; 7:2234–45. [PubMed: 18465893]

44. Ng KK, Petersen JF, Cherney MM, Garen C, et al. Structural basis for the inhibition of porcine pepsin by Ascaris pepsin inhibitor-3. Nat Struct Biol. 2000; 7:653–7. [PubMed: 10932249]

45. Sickmeier M, Hamilton JA, LeGall T, Vacic V, et al. DisProt: the database of disordered Proteins. Nucleic Acids Res. 2007; 35:D786–93. [PubMed: 17145717]

46. Orengo CA, Pearl FM, Bray JE, Todd AE, et al. The CATH database provides insights into protein structure/function relationships. Nucleic Acids Res. 1999; 27:275–9. [PubMed: 9847200]

47. Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol. 2005; 6:197–208. [PubMed: 15738986]

48. Demarest SJ, Martinez-Yamout M, Chung J, Chen H, et al. Mutual synergistic folding in recruitment of CBP/p300 by p160 nuclear receptor coactivators. Nature. 2002; 415:549–53. [PubMed: 11823864]

49. Shaiu WL, Hu T, Hsieh TS. The hydrophilic, protease-sensitive terminal domains of eucaryotic DNA topoisomerases have essential intracellular functions. Pac Symp Biocomput. 1999; 4:578–89. [PubMed: 10380229]

50. Fass D, Bogden CE, Berger JM. Quaternary changes in topoisomerase II may direct orthogonal movement of two DNA strands. Nat Struct Biol. 1999; 6:322–6. [PubMed: 10201398]

51. Bochtler M, Hartmann C, Song HK, Bourenkov GP, et al. The structures of HslU and the ATP-dependent protease HslU-HslV. Nature. 2000; 403:800–5. [PubMed: 10693812]

52. Wang J, Song JJ, Franklin MC, Kamtekar S, et al. Crystal structures of the HslVU peptidase-ATPase complex reveal an ATP-dependent proteolysis mechanism. Structure. 2001; 9:177–84. [PubMed: 11250202]

53. Chen JW, Romero P, Uversky VN, Dunker AK. Conservation of intrinsic disorder in protein domains and families: I. A database of conserved predicted disordered regions. J Proteome Res. 2006; 5:879–87. [PubMed: 16602695]

54. Tompa P, Fuxreiter M, Oldfield CJ, Simon I, et al. Close encounters of the third kind: disordered domains and the interactions of proteins. Bioessays. 2009; 31:328–35. [PubMed: 19260013]

55. Pettersen EF, Goddard TD, Huang CC, Couch GS, et al. UCSF Chimera? A visualization system for exploratory research and analysis. J Comput Chem. 2004; 25:1605–12. [PubMed: 15264254]

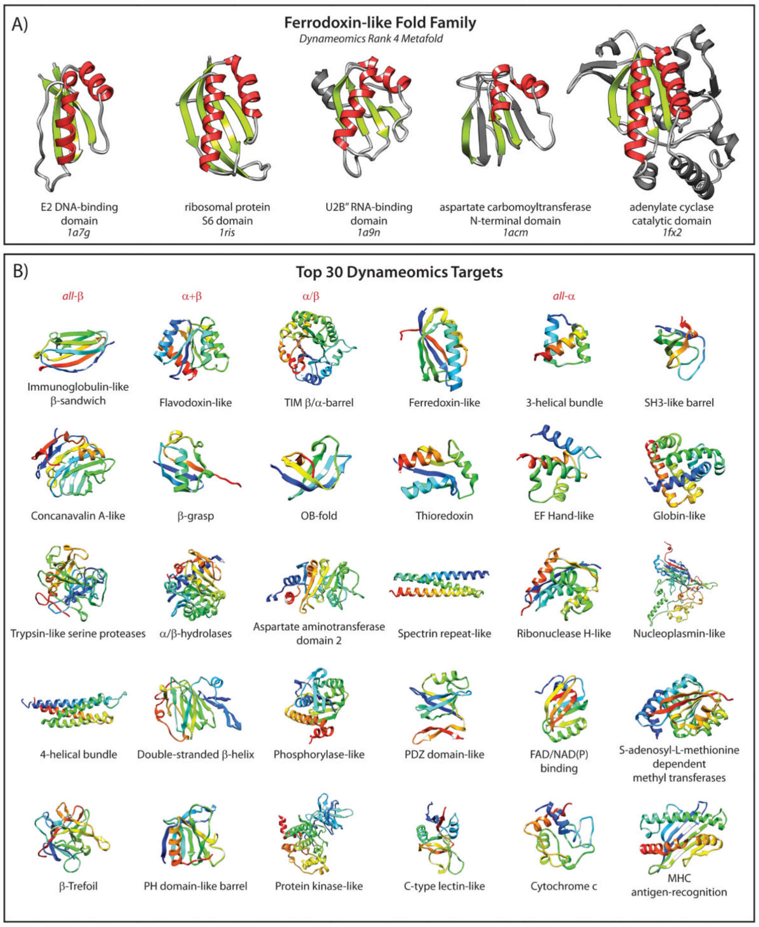

**Figure 1.**

Variations in protein domains within a protein fold family and variations between different fold families. A: The ferredoxin-like fold family with α-helices in red and β-strands in green. The structural elements shared by different members of the family are highlighted, with non-consensus structure in gray. B: Representative structures of the Top 30 most populated fold families from the Dynameomics project colored from N- to C-terminus (blue to red).

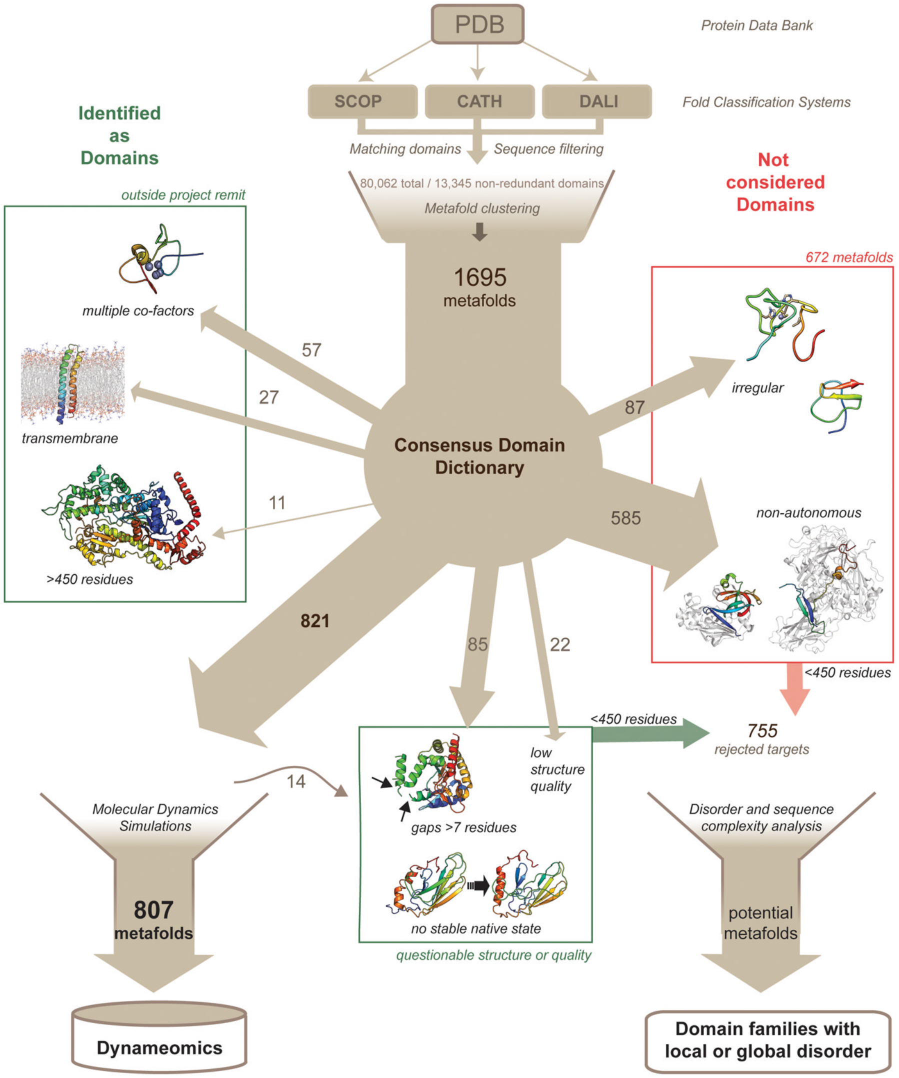

**Figure 2.**
The steps taken in creation of the v2009 CDD and the selection of the Dynameomics targets with the breakdown of the rejected targets detailed. Rejected targets classed as domains but rejected for quality or simulation constraints are boxed in green; those not considered to be domains at that time are boxed in red. The non-autonomous, irregular structures, unstable simulations and crystal structures with coordinate gaps, all less than 450 residues in length, form the set of 755 rejected targets that were surveyed for the possibility of disordered regions.

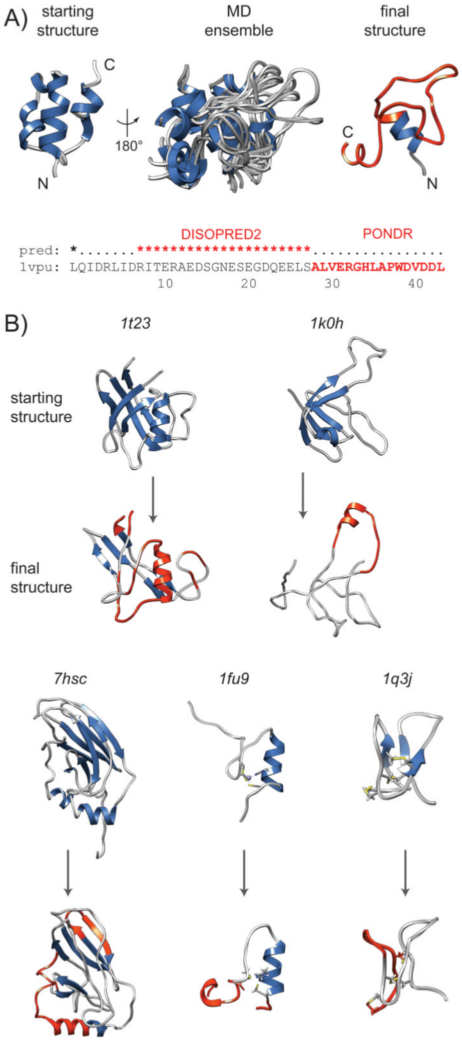

**Figure 3.**

Metafold representatives with putative or confirmed disordered regions that were rejected from the Dynameomics project due to simulation instability. Targets have secondary structure colored blue. The final structures showing the conformational changes post-simulation have regions highlighted in red where disorder was predicted. A: HIV protein Vpu (PDB ID: *1vpu*) that was predicted both here (DISOPRED2) and previously (PONDR) [32] to be substantially disordered, with correlating variation in the 80 ns MD ensemble (10 ns snapshots) and loss of secondary structure in the final structure. Inset is the DISOPRED2 prediction, with previously predicted region also highlighted. B: Starting and final structures of five additional simulations with greater than 10 contiguous residues predicted to be disordered: *1t23; 1k0h; 7hsc; 1fu9; 1q3j.*

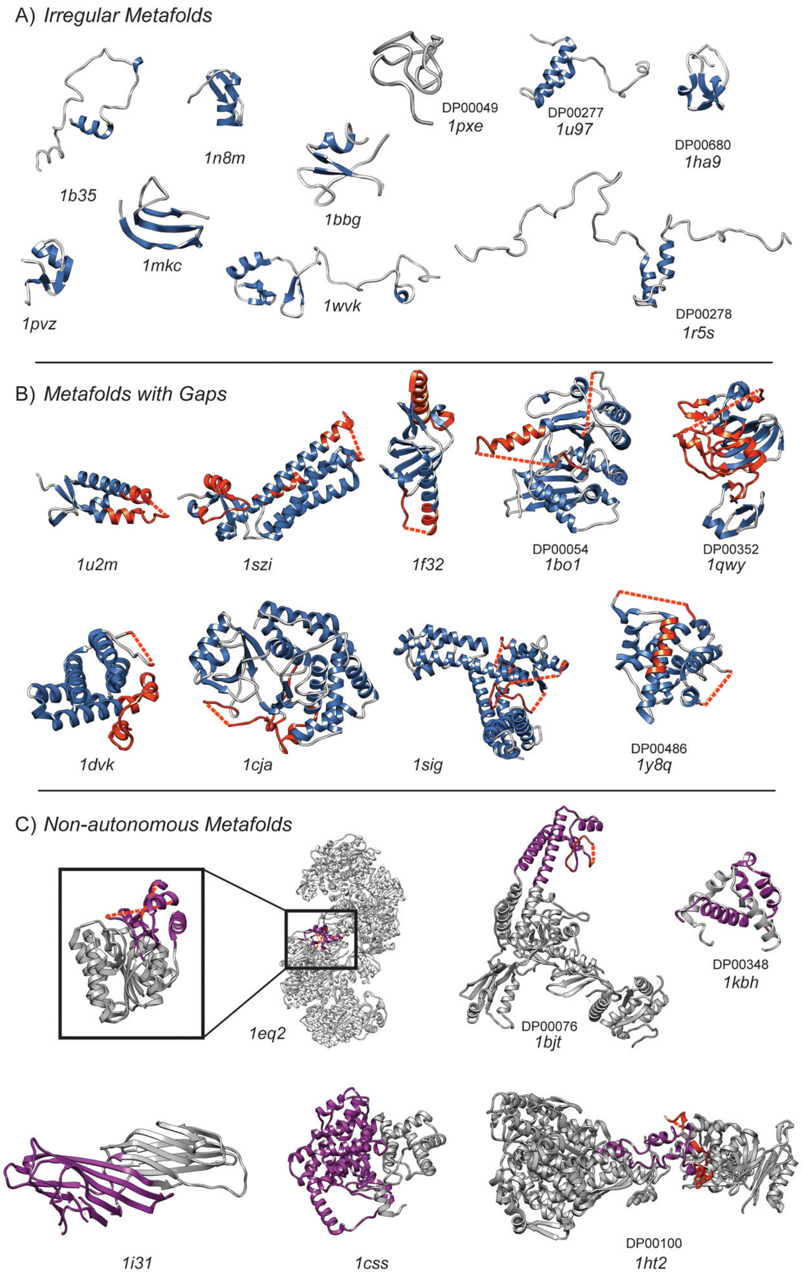

**Figure 4.**

Structures of rejected targets that were initially rejected for being irregular, non-autonomous or containing significant gaps in structures where coordinates could not be experimentally defined. PDB and DisProt codes are inset where membership applies. A: Irregular metafolds with secondary structure colored in blue. Disordered regions mostly coincide with the coil regions colored gray and are not highlighted. B: Metafolds with missing coordinates in X-ray structures, secondary structure is colored blue with disordered regions highlighted in red. Where the disorder pertains to a missing region, the gap is marked with a dashed line. C: Non-autonomous or discontinuous metafolds. The target that is disordered in isolation or incapable of autonomous folding is colored purple, with interrupting structure or extraneous members of a complex shaded in gray. Predicted or experimentally confirmed disordered regions are highlighted in red, where this correlates with missing coordinates a dashed line is used to denote the gap.