

**UCSF**

**UC San Francisco Electronic Theses and Dissertations**

**Title**

A Global View of Structure-function Relationships in the Tautomerase Superfamily

**Permalink**

<https://escholarship.org/uc/item/7h49d6jf>

**Author**

Davidson, Rebecca

**Publication Date**

2018

Peer reviewed|Thesis/dissertation

A Global View of the Sequence-Structure-Function Relationships in  
the Tautomerase Superfamily.

by

Rebecca Davidson

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biological and Medical Informatics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Copyright 2018

by

Rebecca Davidson

To my best friend, Rebeca Salas-Boni, without whose support this would not have been  
accomplished.



## Acknowledgements

This dissertation is the product of collaborative work and that could not have occurred without the support and dedication of several key individuals. First, I'd like to thank my advisor Patricia Babbitt who has provided me with constant support and experiences that have helped me grow. I give her tremendous credit for carefully breeding a lab environment full of brilliant, inquisitive, supportive, unique individuals. In particular, I would like to thank Gemma Holliday who has been a key mentor in the history of my life. I will never forget her kindness and support, and the sheer patience with which she would answer a never-ending barrage of questions. I would also like to thank Eyal Akiva and David Mischel for helping me think critically about my work and incredibly useful advice spanning many challenges. I would also like to thank Elaine Meng and Ben Polacco for their support.

I am especially thankful to Christian P. Whitman and Bert-Jan Baas, who was co-first author on the publication resulting from Chapter 2, for an extremely engaging and fruitful collaboration and for the comedic lightness brought to our regular Skype meetings between Texas and San Francisco.

I am lucky to have an amazing panoply of brilliantly unique friends to have supported me through the past five years. I will again mention again Gemma Holliday and Rebeca Salas-Boni. In San Francisco Kevin Malek, Mimi Nick helped me find my way to the end of my PhD career. From back home in New York, I am grateful for the continual support of Benjy Neymotin, Bryan Meisel, Miriam Rosenberg, Rochelle Furman, and Moin Chowdhury. Finally, I would like to thank my cousin Jodi Gerbi for being my constant cheerleader, always on my side.

# Abstract

Innovating the use of protein similarity networks  
within the context of the Tautomerase Superfamily

Rebecca Davidson

In the genomic age, there is so much data that experimental validation on all of it is impractical. Therefore, computational methods need to be employed to probe this influx of sequence data. Even though we can only experimentally characterize a small amount of these proteins, the sequence data carries a tremendous amount of information through the genetic diversity introduced. This work is concerned with strategies that work towards harnessing sequence similarity networks to their fullest potential within the context of a particular enzyme superfamily, the Tautomerase Superfamily (TSF).

Sequence similarity networks (SSNs), are a type of similarity network in which the nodes represent proteins (or groups of similar proteins as in the representative networks) and the edges represent the similarity between two nodes. The edges are well defined and calculated as the pairwise BLAST *E*-value. These networks provide a graphical view of the similarity relationships within a set of proteins and provide a means to facilitate large-scale analyses. They can also be studied analytically using a variety of algorithms and allow the all-by-all comparisons of tens of thousands of proteins in an intuitively accessible manner.

In this work, a new method of probing SSNs is presented in application to the TSF. This method involves identifying a similarity path through a network with the sequences on the path exhibiting transitional functional features. This path was then used to guide target selection for crystallization of transition linker proteins. Those targets are presented in the context of larger

bioinformatics analysis, including a phylogenetic reconstruction, and an experimental kinetic analysis.

Additionally, the classification of the TSF into subgroups, and in some cases a finer level of clustering, is provided. This curated work complete with sequence alignments and HMMs is hosted on the Structure Function Linkage Database for open access to the scientific community, with the alignments additionally available on GitHub.

# Table of Contents

<b>Chapter 1. Introduction .....</b>	<b>1</b>
<b>1.1 Biology in the genomic era.....</b>	<b>1</b>
<b>1.2 Scope of this dissertation .....</b>	<b>1</b>
<b>Chapter 2. A Global View of Structure-function Relationships in the Tautomerase Superfamily .....</b>	<b>3</b>
<b>2.1 Abstract.....</b>	<b>3</b>
<b>2.2 Introduction .....</b>	<b>4</b>
<b>2.3 Results and Discussion .....</b>	<b>8</b>
2.3.1. A large-scale comparison 11,395 sequences reveals new structural and functional features of the TSF .....	8
2.3.2. Some TSF sequences lack an N-terminal proline .....	12
2.3.3. Phylogenetic representation in the TSF .....	14
2.3.4. Some MIF proteins in higher eukaryotes function as cytokines .....	16
2.3.5. Structure-function relationships among TSF subgroups: 4-OT and <i>cis</i> -CaaD .....	17
2.3.6. Phylogenetic reconstruction of the linker set .....	24
2.3.7. Additional linkers among other TSF subgroups .....	26
<b>2.4 Summary .....</b>	<b>28</b>
<b>2.5 Experimental Procedures .....</b>	<b>29</b>
<b>Chapter 3. Curation of the Tautomerase Superfamily.....</b>	<b>40</b>
<b>3.1 Introduction .....</b>	<b>40</b>
<b>3.1 Level-1 Subgroups .....</b>	<b>42</b>
3.1.1. 4-oxalocrotonate tautomerase .....	42
3.1.2. 5-(carboxymethyl)-2-hydroxymuconate isomerase .....	43
3.1.3. <i>cis</i> -3-chloroacrylic acid dehalogenase subgroup.....	44
3.1.4. macrophage migration inhibitory factor subgroup.....	45
3.1.5. malonate semialdehyde decarboxylase subgroup .....	46

<b>Level-2 Subgroups</b> .....	<b>47</b>
3.1.6. 4-oxalocrotonate tautomerase – Group 1 .....	47
3.1.7. 4-oxalocrotonate tautomerase – Group 2 .....	48
3.1.8. 4-oxalocrotonate tautomerase – Group 3 .....	49
3.1.9. 4-oxalocrotonate tautomerase – Group 4 .....	50
<b>Chapter 4. Conclusion</b> .....	<b>51</b>
<b>References</b> .....	<b>52</b>
<b>Appendix A. Supplemental Information for Chapter 2</b> .....	<b>58</b>
<b>Appendix B. Supplemental Information for Chapter 3</b> .....	<b>78</b>
<b>B.1 Sequence alignments and HMMs for level-1 subgroups</b> .....	<b>78</b>
B.1.1 5-(carboxymethyl)-2-hydroxymuconate isomerase .....	78
B.1.2 cis-3-chloroacrylic acid dehalogenase .....	78
B.1.3 macrophage migration inhibitory factor .....	78
B.1.4 Malonate semialdehyde decarboxylase .....	78
<b>B.2 Sequence alignments and HMMs for level-2 subgroups of the 4-oxalocrotonate tautomerase subgroup</b> .....	<b>79</b>
B.2.1 Group 1 .....	79
B.2.2 Group 2 .....	79
B.2.3 Group 3 .....	79
B.2.4 Group 4 .....	79

## List of Figures

Figure 2.1 Major types of characterized reactions in the TSF .....	6
Figure 2.2 Representative sequence similarity network of the TSF superfamily .....	9
Figure 2.3 Representative SSN showing sequences that lack an N-terminal proline .....	13
Figure 2.4 Phylogenetic representation in the TSF.....	15
Figure 2.5 Sequences in the 50% representative network that link the 4-OT and <i>cis</i> -CaaD subgroups.....	19
Figure 2.6 Structural comparison of conserved active site residues in Linker 1 and Linker 2.....	23
Figure 2.7 Phylogenetic tree of the 4-OT and <i>cis</i> -CaaD linkers from the 50% representative SSN..	25
Figure 2.8 Structure similarity network of the TSF .....	27
Figure 3.1 The Tautomerase superfamily in the SFLD.....	41
Figure 3.2 SFLD page for the 4-oxalocrotonate tautomerase subgroup.....	42
Figure 3.3 SFLD page for the 5-(carboxymethyl)-2-hydroxymuconate isomerase subgroup .....	43
Figure 3.4 SFLD page for the <i>cis</i> -3-chloroacrylic acid dehalogenase subgroup .....	44
Figure 3.5 SFLD page for the macrophage migration inhibitory factor subgroup .....	45
Figure 3.6 SFLD page for the malonate semialdehyde decarboxylase subgroup. ....	46
Figure 3.7 SFLD page for Group 1 of the Oxalocrotonate Tautomerase Subgroup .....	47
Figure 3.8 SFLD page for Group 2 of the Oxalocrotonate Tautomerase Subgroup .....	48
Figure 3.9 SFLD page for Group 4 of the Oxalocrotonate Tautomerase Subgroup .....	49
Figure 3.10 SFLD page for Group 4 of the Oxalocrotonate Tautomerase Subgroup .....	50
Figure A.1 Oligomeric organization in the TSF.....	58
Figure A.2 Mapping of the Level 1 <i>cis</i> -CaaD HMM to the Level 1 <i>cis</i> -CaaD subgroup.....	59
Figure A.3 Mapping of the Level 1 MSAD HMM to the Level 1 MSAD subgroup .....	60
Figure A.4 Mapping of the Level 1 CHMI HMM to the Level 1 CHMI subgroup .....	61
Figure A.5 Mapping of the Level 1 MIF HMM to the Level 1 MIF subgroup. ....	62

<b>Figure A.6 90% sequence identity per node network of Level 2 subgroups of the Level 1 4-OT subgroup .....</b>	<b>63</b>
<b>Figure A.7 HMM mapping of the Level 2 subgroup 1 to the Level 1 4-OT subgroup.....</b>	<b>64</b>
<b>Figure A.8 HMM mapping of the Level 2 subgroup 2 to the Level 1 4-OT subgroup.....</b>	<b>65</b>
<b>Figure A.9 HMM mapping of the Level 2 subgroup 3 to the Level 1 4-OT subgroup.....</b>	<b>66</b>
<b>Figure A.10 HMM mapping of the Level 2 subgroup 4 to the Level 1 4-OT subgroup.....</b>	<b>67</b>
<b>Figure A.11 Length histogram of 11,395 non-redundant protein sequences in the TSF .....</b>	<b>68</b>
<b>Figure A.12 Non-Pro-1 Frequencies in the TSF.....</b>	<b>70</b>
<b>Figure A.14 MSA of sequences used to calculate the phylogenetic tree .....</b>	<b>71</b>
<b>Figure A.15 Examples of the curation process used to validate the non-Pro-1 sequences in the TSF .....</b>	<b>74</b>
<b>Figure A.16 PDB codes of structures used in the structure similarity network provided in Figure 2.8.....</b>	<b>76</b>

## List of Tables

Table 2.1 Tautomerase activity across the linker proteins .....	22
---	----



# **Chapter 1. Introduction**

## **1.1 Biology in the genomic era**

As the already tremendous amounts of genomic data continue to grow, experimental validation on all of it becomes increasingly impractical. As a result computational analysis of the vast amounts of sequence data being produced becomes more and more important. Naturally, developing and expanding methods that analyze this data and help biologists form hypotheses to guide experimentation becomes an enticing challenge in this age of genomics and data science.

A computational approach that has traditionally been employed in the Babbitt lab to study the sequence data of large mechanistically diverse enzyme superfamilies is Sequence Similarity Networks (SSNs). These networks provide a graphical view of the similarity relationships within a set of proteins and provide a means to facilitate large-scale analyses. Biologists can use these to elucidate trends in sequence function relationships and guide making hypotheses to plan experiments while seeing how any particular hypothesis relates to the larger set of proteins in sequence space. The Babbitt lab pioneered methods for the large-scale classification and discovery of function in enzyme superfamilies using these SSNs. In addition to studying Superfamilies, finding new and creative ways to probe hypotheses within the SSN framework is another key goal of the Babbitt lab.

## **1.2 Scope of this dissertation**

In this work, we use a new method of probing SSNs that involves identifying a similarity path through a network with the sequences on that path exhibiting transitional functional features. Our analysis of the tautomerase superfamily using these methods has been published (1) and the

accepted paper is provided here as Chapter 2 of this thesis. Importantly, this path was used to guide target selection for crystallization of transitional linker proteins, that contributed to a structure guided comparison. Beyond that, this path and the resulting structure guided alignment were used to guide sequence selection for a phylogenetic reconstruction. Both the kinetic analysis of the linker proteins and the phylogenetic reconstruction corroborated the proposed structure/function transitions, exhibiting the utility of this approach. This new linker guided strategy can be employed in the future study of this and other superfamilies to probe sequence/structure/function relationships.

Accompanying the work described in Chapter 2, Chapter 3 describes our new classification among the proteins of Tautomerase superfamily. This was done by breaking them in subgroups, and in some cases finer groups within the subgroups. The classification suggests a finer granularity of functional division among the proteins than suggested by INTERPRO and PFAM. This work is curated and hosted on the Structure Function Linkage Database (2) (<http://sfld.rbvi.ucsf.edu/django/superfamily/159/>) for open access for the scientific community.

## Chapter 2. A Global View of Structure-function Relationships in the Tautomerase Superfamily

### 2.1 Abstract

The tautomerase superfamily (TSF) consists of more than 11,000 non-redundant sequences present throughout the biosphere. Characterized members have attracted much attention because of the unusual and key catalytic role of an N-terminal proline. These few characterized members catalyze a diverse range of chemical reactions, but the full scale of their chemical capabilities and biological functions remains unknown. To gain new insight into TSF structure–function relationships, we performed a global analysis of similarities across the entire superfamily and computed a sequence-similarity network to guide classification into distinct subgroups. Our results indicated that TSF members are found in all domains of life, with most being present in bacteria. The eukaryotic members of the *cis*-3-chloroacrylic acid dehalogenase subgroup are limited to fungal species, while the macrophage-migration inhibitory factor subgroup has wide eukaryotic representation (including mammals). Unexpectedly, we found that 346 TSF sequences lack Pro-1, of which 85% are present in the malonate semialdehyde decarboxylase subgroup. The computed network also enabled identification of similarity paths, namely sequences that link functionally diverse subgroups and exhibit transitional structural features that may help explain reaction divergence. A structure-guided comparison of these linker proteins identified conserved transitions between them, and kinetic analysis paralleled these observations. Phylogenetic reconstruction of the linker set was consistent with these findings. Our results also suggest that contemporary TSF members may have evolved from a short 4-oxalocrotonate tautomerase–like ancestor, followed by gene duplication and fusion. Our

new linker-guided strategy can be used to enrich discovery of sequence/structure/function transitions in other enzyme superfamilies.

## **2.2 Introduction**

As the number of protein sequences from genomic and metagenomic sequencing continues to grow exponentially, the proportion of these sequences accessible to experimental function determination becomes vanishingly small, even using high throughput approaches. For enzymes, clues about reaction diversity that has evolved across the biosphere have contributed significantly to the development of principled and generalizable approaches for predicting functional capabilities of proteins of unknown function (“unknowns”) and for using that information to identify informative targets for biochemical characterization or protein engineering. Mechanistically diverse enzyme superfamilies (3) (also called functionally diverse superfamilies) represent about a third of the universe of enzyme superfamilies (4). These superfamilies often contain more than 20,000 homologous sequences in which active site machinery associated with a conserved aspect of catalysis is maintained through evolution while specialized active site and other structural variations evolve to enable many different reactions. Studying these systems offers powerful insight into ways that nature has produced the enormously varied reactions necessary for life and provides a framework for inferring their functions. A small number of these superfamilies have now been studied on a large scale, revealing for each the structural and mechanistic foundations by which divergent evolution has produced many different reactions.

Here, we describe a global mapping of structure-function relationships among the members of the tautomerase superfamily (TSF)<sup>1</sup> (5, 6). Although the TSF has not yet been examined on a large scale, its few known reactions reflect the presence of a treasure trove of enzymes with unusual properties. Of special interest, the unusual mechanistic use of an N-terminal proline offers a window into nature's use of "outlier" catalytic strategies (5, 7), broadening our still limited understanding of the chemistry supported by functionally diverse enzyme superfamilies. Especially important for the work described here, the relatively simple structural organization of TSF proteins has allowed us to uncover sequence variations among contemporary TSF sequences that may provide clues about how a simple ancestral scaffold may have diverged to produce widely varied reaction types and biological functions. TSF members are highly desirable experimental vehicles as they do not require metal ions or coenzymes and are easily purified and expressed (6). As a result, computational predictions can be experimentally tested, enabling an iterative strategy for choosing and testing unknowns likely to inform a better understanding of structure, function, and mechanism across the superfamily.

All biochemically characterized enzymes in the superfamily fall into five reaction types (Figure 2.1), each of which uses the N-terminal proline either as a general base or a general acid (6, 8). All previous studies of characterized TSF members show a shared utilization of an N-terminal proline in their mechanisms, leading to an expectation that all TSF members would exhibit this feature. Notably, the common reaction catalyzed by three of these enzymes, 4-oxalocrotonate tautomerase (4-OT) (7, 9), 5-(carboxymethyl)-2-hydroxymuconate isomerase

---

<sup>1</sup> The abbreviations used are: CHMI, 5-(carboxymethyl)-2-hydroxymuconate isomerase; *cis*-CaaD and CaaD, *cis*- and *trans*-3-chloroacrylic acid dehalogenase, respectively; CgX, Cg10062 from *Corynebacterium glutamicum*; D-DT, D-dopachrome tautomerase; f4-OT, fused 4-OT; hh4-OT, heterohexamer 4-OT; HMM, hidden Markov model; 2-HM, 2-hydroxymuconate; MIF, macrophage migration inhibitory factor; MSAD, malonate semialdehyde decarboxylase; MSA, multiple sequence alignment; 4-OT, 4-oxalocrotonate tautomerase; PP, phenylpyruvate; PPT, phenylpyruvate tautomerase; PDB, protein data bank; SSN, sequence similarity network; SFLD, Structure-Function Linkage Database; TSF, tautomerase superfamily; *wb*, *Wuchereria bancrofti*.

(CHMI) (10), and the phenylpyruvate tautomerase (PPT) activity of macrophage migration inhibitory factor (MIF) (11-13) is an enol-keto tautomerization of a pyruvoyl moiety in which the Pro-1 has a low  $pK_a$  value ( $\sim 6.4$  in 4-OT) (9). The two other reaction types, catalyzed by *cis*- and *trans*-3-chloroacrylic acid dehalogenase (*cis*-CaaD and CaaD, respectively) (6, 14-16) and malonate semialdehyde decarboxylase (MSAD) (17), are more divergent in mechanism. While these latter two enzyme-catalyzed reactions still utilize the N-terminal proline, it has a higher  $pK_a$  value ( $\sim 9.2$  in CaaD) (18) and functions as a general acid (17). In addition to its PPT activity, MIF is more commonly recognized as functioning as a pro-inflammatory cytokine in mammals (19-22). The PPT and the thiol-protein oxidoreductase activities (23, 24) of MIF are not involved in its cytokine activities (25, 26).

Enzyme	Monomer length	Oligomeric state	Reaction
4-OT (EC 5.3.2.6)	62	homohexamer	<p>2-hydroxymuconate <math>\rightleftharpoons</math> 2-oxohex-3-enedioate</p>
CHMI (EC 5.3.3.10)	125	trimer	<p>5-(carboxymethyl)-2-hydroxymuconate <math>\rightleftharpoons</math> 5-(carboxymethyl)-2-oxohex-3-enedioate</p>
MIF (EC 5.3.2.1)	114	trimer	<p>phenylenolpyruvate <math>\rightleftharpoons</math> phenylpyruvate</p>
<i>cis</i> -CaaD (EC 3.8.1.-)	149	trimer	<p><i>cis</i>-3-chloroacrylate <math>\xrightarrow{+H_2O/-HCl}</math> malonate semialdehyde</p>
MSAD (EC 4.1.1.-)	129	trimer	<p>malonate semialdehyde <math>\longrightarrow</math> acetaldehyde + CO<sub>2</sub></p>

**Figure 2.1 Major types of characterized reactions in the TSF.** 4-OT (26); CHMI: (10), PPT activity of MIF: (13), *cis*-CaaD: (27), and MSAD: (28). The pyruvoyl-moiety that is the common functional group for the three tautomerase reactions (of 4-OT, CHMI and MIF) is boxed in red for the 4-OT reaction. The proton that is transferred during each of these reactions is highlighted in red.

From a structural perspective, the functionally diverse members of the TSF share a small, structurally similar core domain comprised of a  $\beta$ - $\alpha$ - $\beta$  motif. The smallest members of the TSF are comprised of monomers ranging in length from 58-84 amino acids. For the handful of nonredundant structures available, oligomeric organization across the superfamily varies considerably and includes homo- or heterohexamers (a single  $\beta$ - $\alpha$ - $\beta$  unit per monomer) (29, 30), trimers (two fused  $\beta$ - $\alpha$ - $\beta$  units) (13, 27, 28), and a dimer (a single  $\beta$ - $\alpha$ - $\beta$  unit) (31) (See Supplemental Figure 1 for details). Both observed and as yet uncharacterized variations among oligomeric organization may impact the *in vivo* catalytic capabilities of these proteins and contribute to expansion of the functional repertoire of the TSF as well. Although most TSF members appear to be composed of either the core  $\beta$ - $\alpha$ - $\beta$  motif or the fused duple form, the motif is also found appended to a limited number of non-ribosomal peptide synthetases such as indigoidine synthetase (32). For clarity of discussion, the short TSF members constructed of a single  $\beta$ - $\alpha$ - $\beta$  structural motif will be denoted in this work as a single  $\beta$ - $\alpha$ - $\beta$  domain, and the longer TSF members consisting of two fused  $\beta$ - $\alpha$ - $\beta$  structural motifs will be denoted as two  $\beta$ - $\alpha$ - $\beta$  subdomains.

Despite the attention that individual enzymes in the TSF have received, the full extent of the chemical and structural diversity across the superfamily remains unknown, as fewer than 30 members have been experimentally characterized. As with other functionally diverse superfamilies, this lack of experimental data prompts many questions. Do all TSF members have a catalytic Pro-1? Is the breadth of the superfamily reaction space confined to the five known reaction classes? Has nature used other arrangements for the core  $\beta$ - $\alpha$ - $\beta$  motif in evolving the contemporary members of the TSF? How did the various subgroups evolve from what has been proposed to be the ancestral 4-OT-like enzyme (5, 24, 33, 34)? How did the multiple biological

activities of MIF evolve? Is there a biological relevance for the reported enzymatic activities of MIF (25), such as the PPT activity (11-13)?

To begin to address some of these questions, we performed an all-by-all comparison of more than 11,000 non-redundant sequences of the TSF, using sequence similarity networks to map key structure-function relationships among the superfamily members. The global views that result begin to fill out the picture of structure-function relationships across the entire TSF and reveal new trends not previously available from small scale studies. The results allow classification of the TSF into subgroups that exhibit different patterns of catalytic machinery and provide a map of its phylogenetic representation across the biosphere. We also address in this work how changes in reaction specificity-determining residues may have led to the evolution of varied functions. Phylogenetic reconstruction focused on enzymes that show sequence similarity to two different subgroups (“linkers”), together with the other results reported here, supports the hypothesis that the *cis*-CaaD-like enzymes may have evolved from a 4-OT-like ancestor.

Similarity networks from this paper, and other data, including multiple sequence alignments (MSAs) resulting from this study are available for download from the Structure-Function Linkage Database (SFLD) (2) (<http://sfld.rbvi.ucsf.edu/django/superfamily/159/>).

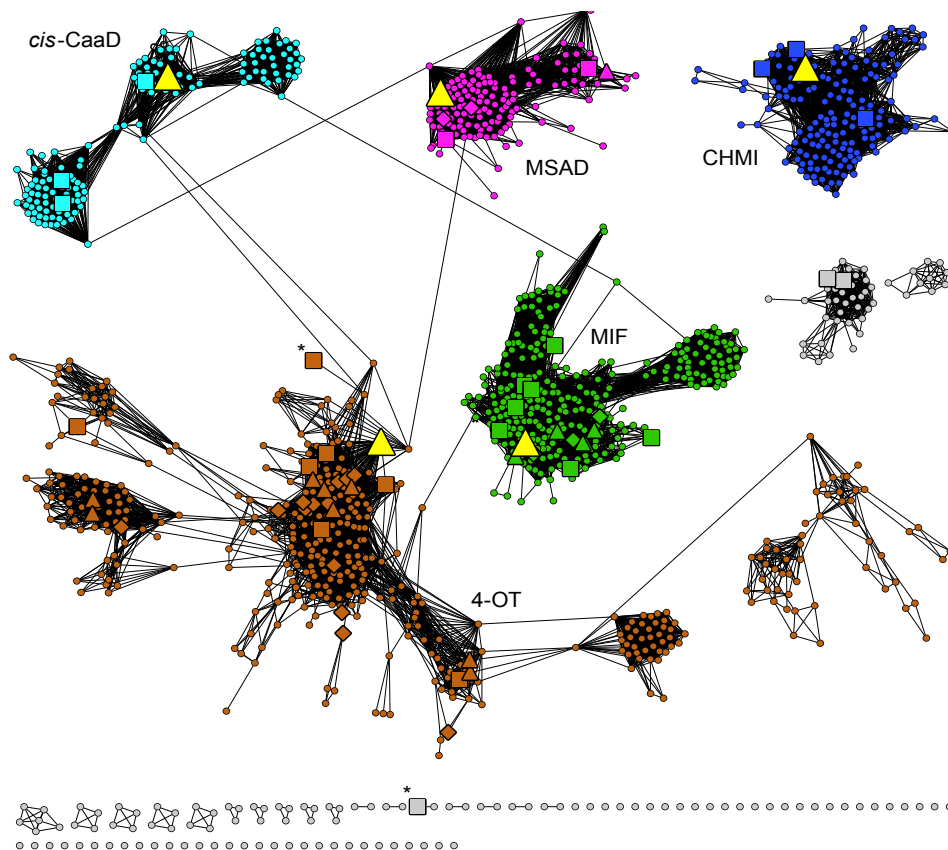
## **2.3 Results and Discussion**

### **2.3.1. A large-scale comparison 11,395 sequences reveals new structural and functional features of the TSF**

All-by-all pairwise comparisons of 11,395 non-redundant sequences of the TSF were computed and used to generate a representative sequence similarity network (SSN) (35, 36) (Figure 2.2). Each node in this network includes all TSF members that share >50% identity. Thus, the number of sequences comprising each node shown in the figure can be highly variable,



ranging from 1-342 sequences per node. Using this network as an initial guide, the sequences were classified into Level 1 subgroups in which the sequences within each are more similar to each other than to the sequences in any other subgroup.



**Figure 2.2 Representative sequence similarity network of the TSF superfamily summarizes putative sequence-function relationships.**

As indicated in the Figure, the majority of TSF sequences can be assigned to a Level 1 subgroup. Although the subgroups were defined based only on their sequence similarities, mapping of the five known enzyme-catalyzed reactions shown in Figure 2.1 to the network indicates that each uniquely belongs to a different subgroup. Based on this observation, the best-characterized protein from each subgroup was termed a “founder” protein and was highlighted in

Figure 2.2. Each subgroup with a founder sequence was assigned a distinct color; other smaller sets of representative sequences and singleton sequences were not named or investigated further.

To provide a quantitative determination of the uniqueness of the subgroupings shown in Figure 2.2, hidden Markov models (HMMs) (37) were computed for each named subgroup of the TSF and searched against all superfamily members to evaluate the degree to which each uniquely describes the subgroup from which it was computed (Supplemental Figures 2-5). The results of this control experiment show that the HMMs generated for the Level 1 *cis*-CaaD, MSAD, CHMI, and MIF subgroups recovered the sequences of their associated subgroups at high scores as expected, with only a few outliers in other subgroups.

As the diversity of the sequences of the 4-OT subgroup was too great to allow generation of a high confidence multiple sequence alignment (MSA) for the entire subgroup, it was further subgrouped into four Level 2 subgroups (Supplemental Figure 6) for which alignments and HMMs could be reasonably generated. As with the main Level 1 subgroups, the Level 2 HMMs for the 4-OTs were mapped to their associated subgroups (Supplemental Figures 2.7-2.10). The results show that these Level 2 subgroup HMMs capture the majority of the subgroup members from which each was generated, although these HMMs have a larger proportion of cross hits to other Level 2 subgroups of the 4-OTs than did the HMMs generated to the Level 1 subgroups. This is likely due to the greater diversity within the 4-OT subgroup relative to the diversity within each of the other Level 1 subgroups of the superfamily.

As has been reported for other functionally diverse enzyme superfamilies (38-41), only a small proportion of TSF members has been biochemically or structurally characterized, as shown in Figure 2.2. Note that only one functionally annotated sequence is required to depict a representative node as a large triangle (indicating an experimentally assigned function) even

though that node may contain a large majority of uncharacterized sequences. This scarcity of functional information for most of the TSF subgroups represents a major barrier to understanding the range of contemporary reactions (or biological functions) the TSF scaffold has evolved to support, limiting, in turn, knowledge that could be useful for re-engineering TSF members to catalyze new reactions in the laboratory (42).

#### *2.3.1.1 Subgroupings may be useful for classifying newly discovered TSF members*

Studies of experimentally characterized 4-OT Level 1 subgroup members have identified key active site residues and patterns for some of the known reactions (7, 9, 10, 29, 43-46). These observations may be useful in predicting functions for uncharacterized TSF members sharing these active site patterns. Likewise, assignment of newly discovered TSF members into the most appropriate subgroups can be achieved using the unique HMMs generated for each subgroup.

At the same time, many of these expected active site patterns that are missing or differ among uncharacterized TSF sequences (data not shown) may suggest that they do not catalyze known reactions of the superfamily, but may have other molecular functions instead. Mapping of more detailed active site variations or motifs to these networks may identify targets for experimental or structural characterization that could shed light about as yet undiscovered reactions of the TSF.

#### *2.3.1.2 Domain structure and sequence length variation across the TSF*

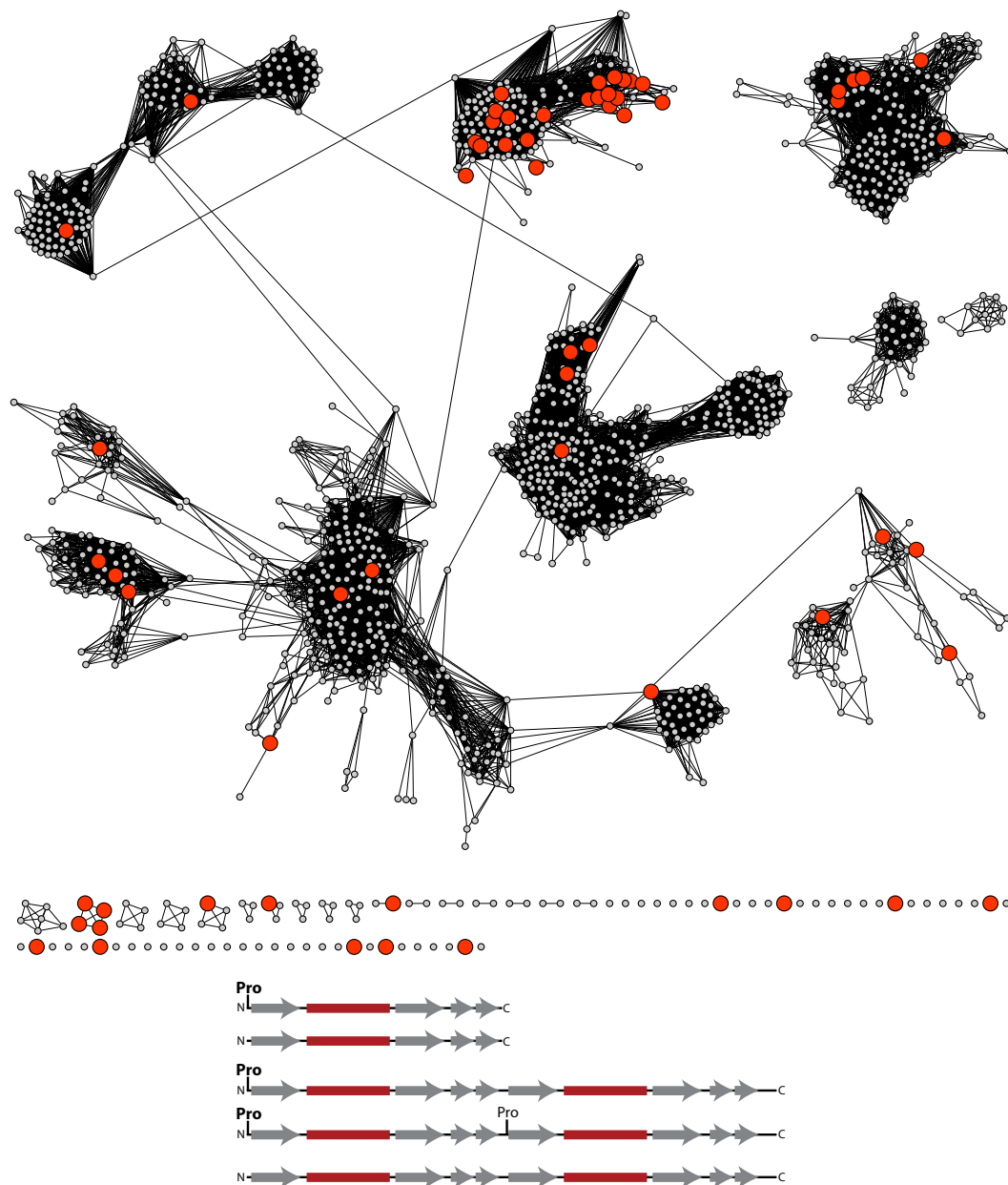
The domain structure of the TSF is relatively simple with nearly all of the members represented by either a single  $\beta$ - $\alpha$ - $\beta$  structural motif or as fusion proteins composed of two of these subdomains. Our results confirm this pattern generally across the TSF; most sequences of the 4-OT subgroup are composed of a short monomer of 58-84 residues except for a small group of fused 4-OTs, discussed below. The great majority of sequences assigned to the other four

subgroups represent the fused form, which are roughly twice as long as the monomeric 4-OTs and are primarily comprised of two fused  $\beta$ - $\alpha$ - $\beta$  subdomains (Supplemental Figure 11). For the longer fused superfamily members, the N- and C-terminal subdomains have diverged to produce unique sequence differences. The remaining superfamily sequences are of mixed size between the short and long proteins, largely due to variable sequence lengths in the regions that link the two  $\beta$ - $\alpha$ - $\beta$  subdomains and at the C-termini. As discussed in more detail in section 2.3.5.1 entitled “*Linkers*” between *cis-CaaD* and 4-OT subgroup identify a similarity path between them, the shorter 4-OT subgroup proteins align better with the C-terminal “half” of the longer fused proteins linking the *cis-CaaD* and 4-OT subgroups than they do with the N-terminal half. The domain structure of the fused proteins of the CHMI, MIF, *cis-CaaD*, and MSAD subgroups may indicate that they evolved by gene duplication and fusion from a simpler 4-OT-like ancestor.

### **2.3.2. Some TSF sequences lack an N-terminal proline**

Experimental work to date has suggested that the hallmark of the TSF, the catalytic N-terminal proline, is a mechanistic imperative. Surprisingly, our global analysis reveals that a significant number (346) of the TSF sequences lack Pro-1 (Figure 2.3, top) (See section 2.5 Experimental Procedures for a description of how this finding was verified.) Although missing this key residue, these sequences are clearly part of the TSF as they align well with other superfamily members and retain conservation of other important active site residues of the subgroup in which they are found. As none of these proteins have been biochemically characterized, their functions remain unknown. The majority of the non-Pro-1 superfamily members (294) map to the MSAD subgroup. Of the remaining 52 sequences, 13 map to the 4-OT subgroup, and the remaining sequences are randomly scattered across the entire network.

Interestingly, the residues that are found in place of Pro-1 are not random variations. Serine, isoleucine, and alanine are the most abundant residues found that follow the initiating methionine (Supplemental Figure 12).



**Figure 2.3 Representative SSN showing sequences that lack an N-terminal proline.** Top: SSN as in Figure 2.2, except that nodes containing one or more sequences that lack an N-terminal proline are colored red. Bottom: Observed positions of the N-terminal proline (or its absence) in sequences of short and fused TSF members. Gray arrows and maroon blocks designate beta strands and alpha helices, respectively.

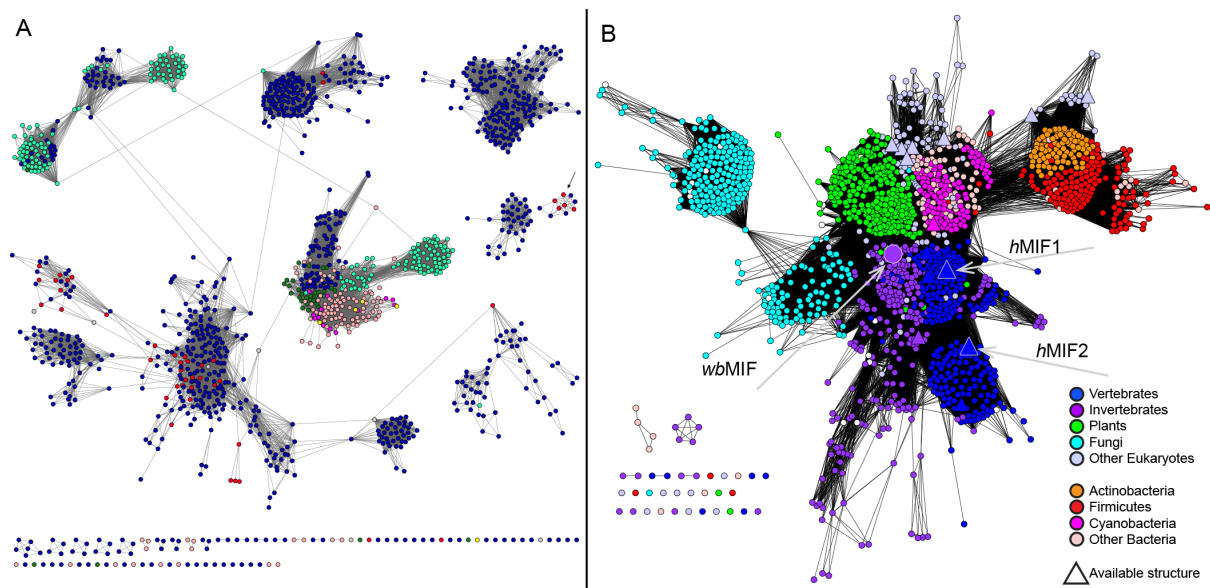
The absence of Pro-1 could reflect functional and/or mechanistic diversity in several ways. It is possible that an N-terminal amino group could function as a general base or acid. The experimentally characterized enzymes of the MSAD group, for example, which contains 85% of the non-Pro-1 proteins, have been shown to require a general acid in the enzyme-catalyzed reactions (17, 28, 47, 48). Or, these proteins could perform catalysis via an as yet uncharacterized alternative catalytic mechanism that doesn't rely on the N-terminal group. Alternatively, some of these proteins may not function as enzymes, but rather, may perform other biological functions, for example, as regulatory proteins. A mechanistic interpretation of these observations is complicated by how the possible forms of non-Pro-1 TSF proteins might be utilized in the varied oligomeric forms currently known (or yet to be identified) in the superfamily (Figure 2.3 and Supplemental Figure 1). An additional complicating issue is that some sequences will retain the initiating methionine and others will not (49). Clearly, experimental characterization of some of these non-Pro-1 sequences will be required to address these and other mechanistic questions.

The second proline in line 4 of the cartoon shown in Figure 2.3 (bottom) provides some support for the notion that the Fused 4-OT members may have evolved by gene duplication and fusion from a short 4-OT ancestor as this proline is conserved in ~60% of the fused members most similar to the founder 4-OT. There are 91 of these Fused 4-OTs and, based on a MSA of the 4-OT Level 2 subgroup 1 (data not shown), 57 likely have proline at the start of the second  $\beta$ - $\alpha$ - $\beta$  subdomain.

### **2.3.3. Phylogenetic representation in the TSF**

As with many large superfamilies, members of the TSF are found across the three domains of life. As shown in Figure 2.4A, bacteria dominate in most of the TSF Level 1

subgroups, with a somewhat lesser relative representation in the *cis*-CaaD and MIF subgroups. While a small set of 14 archaeal sequences is found in the MSAD subgroup, the 4-OT subgroup is significantly enriched in archaeal sequences (139 sequences representing 71% of the 197 archaeal sequences in the TSF), suggesting an ancient origin for some of these 4-OT subgroup proteins. All of these sequences are uncharacterized except one, although its physiological function remains unknown (50). The small sequence cluster not named as a subgroup (marked by the arrow) contains a substantial proportion of archaeal sequences as well. TSF members from eukaryotes are largely confined to the *cis*-CaaD subgroup, which is principally composed of proteins from fungi, and to the MIF subgroup. In the MIFs, the preponderance of proteins comes from eukaryotic phyla, and includes substantial representation from fungi, plants, invertebrates, and vertebrates, including mammals. Although this subgroup includes significant representation from bacterial organisms, it does not appear to include sequences from Archaea.



**Figure 2.4 Phylogenetic representation in the TSF.** A. Network as in Figure 2.2, except that the representative nodes are colored by dominant type of life: red, archaea; dark blue, bacteria; green, plants; cyan, fungi; pink, invertebrates; yellow, mammals; magenta, other vertebrates besides mammals; 26 gray nodes scattered among the dark blue nodes of the 4-OT subgroup come from environmental sequencing projects. The arrow marks an unnamed

subgroup enriched in archaeal sequences. **B.** A one-sequence per node similarity network of the MIF subgroup. Network nodes represent 1,679 MIF sequences, 273 of which come from vertebrates. Coloring is by type of life as shown in the key except for white nodes, which were not designated by type of life in the UniProt database. Triangles represent proteins with solved structures. Large triangles represent two characterized human proteins, MIF1 (Uniprot P14174) and MIF2 (Uniprot P30046) and the large circle represents a MIF from the human parasite *Wuchereria bancrofti* (Uniprot O44786) as indicated by the arrows. The threshold for drawing edges between each node is  $10e^{-18}$ . The network shows two separate MIF groups in vertebrates; the sequences of the group containing the human MIF1 sequence are more similar to their invertebrate relatives than to the group containing the human MIF2 sequence.

### 2.3.4. Some MIF proteins in higher eukaryotes function as cytokines

Figure 2.4B provides greater detail regarding how MIF sequences from varied organisms relate to each other. Among the MIFs, a total of 10 human MIF-like proteins are included in the network. Some of these may represent different splicing variants of MIF1 and of a second group of proteins known as MIF2, (previously known as D-dopachrome tautomerase (D-DT) (51)). Vertebrates represent 273 (16%) of the sequences in this SSN. Within the subgroup, a large set of proteins, including the human MIF1, is principally associated with cytokine activities (22), along with some enzymatic activities (i.e., PPT and protein-thiol oxidoreductase activities) (23-25). A second group of similar size includes the human MIF2 (52). The human MIF1 and MIF2 structures are similar and their subunit topologies are almost identical although there are differences in the active site regions (21). Additionally, the functional properties of MIF2 overlap with those of MIF1, and the two may act cooperatively (53). MIF2 exhibits a tautomerase activity that requires Pro-1 (51) using D-dopachrome, a compound with unknown biological relevance. Its relative tautomerase activity is about 10-fold less than that of MIF1 using the same compound. The MIF2-catalyzed tautomerization is followed by a decarboxylation reaction (25, 53). (The observation that tautomerization of L-dopachrome is involved in melanotic encapsulation, a hallmark of a primitive invertebrate defense pathway, has suggested that the D-DT activity might be a vestigial property (53).) Interestingly, analysis of the network shown in Figure 2.4B at more stringent *E*-value cutoffs for drawing edges (not shown) confirms



the inference suggested in Figure 2.4B, *i.e.*, that the MIF1 proteins are more closely related to the invertebrate members of the subgroup than they are to the MIF2 proteins. Some of these invertebrate organisms are pathogenic (25).

A recent phylogenetic analysis of MIF homologs in plants revealed that an ancestral MIF-like sequence was present in the last common eukaryotic ancestor before plants diverged about 1.6 billion years ago (54) and notes that MIF-like proteins are found in many phyla, including bacterial ones. The large-scale view of MIF similarity relationships illustrated in Figure 2.4B is consistent with those findings and indicates the relative proportion of MIF-like proteins found in bacteria, plants, and fungi. Many of these other organisms harboring MIF-like proteins do not have cytokines or the types of complex immune systems typically associated with cytokine activity and thus may have as yet unknown functional roles. Another published phylogenetic reconstruction of experimentally studied MIFs notes that they may have evolved from an ancient defense molecule; in parasites MIF-like proteins appear to function as virulence factors (25). A MIF-like protein has been biochemically characterized from the human parasite *Wuchereria bancrofti* (wbMIF) and shows a low level of tautomerase activity (55). This latter analysis is the only attempt to assign function to non-mammalian MIFs. Thus, both the molecular and biological functions and the importance of the broad representation of the MIFs across the biosphere remain poorly understood.

### **2.3.5. Structure-function relationships among TSF subgroups: 4-OT and *cis*-CaaD**

The results of this global study of the TSF show that all members of the functionally diverse TSF are united by the structure similarities of a common fold and conservation in key aspects of their chemistry and the catalytic machinery (with the exception of the 3% that lack an N-terminal proline). At the same time, the known TSF reactions are different from each other,

raising the question how the fundamental structure-function relationship that defines the TSF has been modified by nature to evolve these different overall reactions.

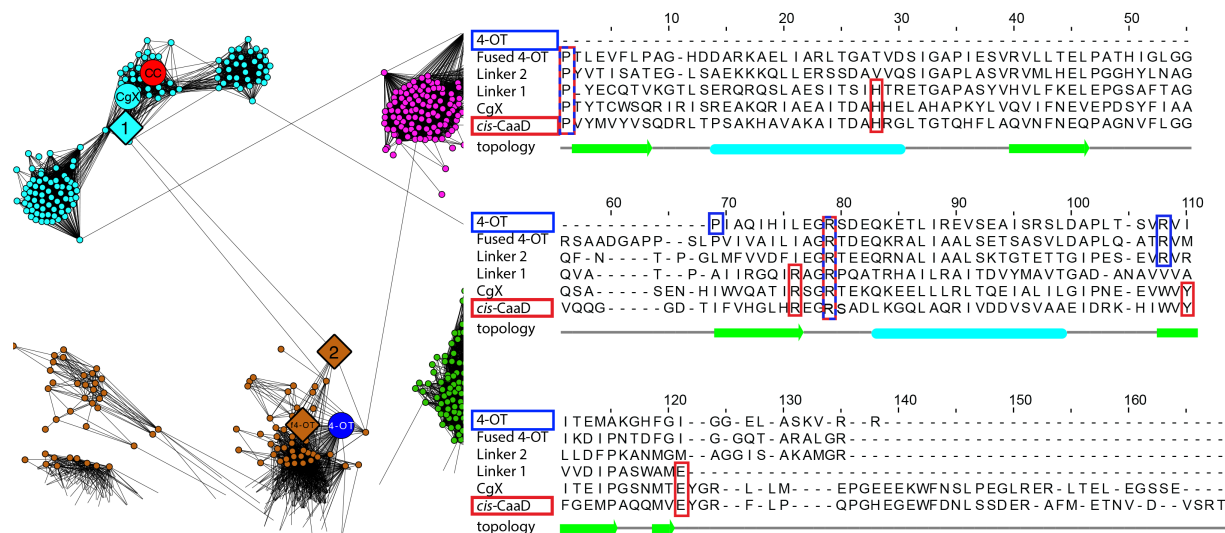
Several issues make answering this question especially difficult. First, the extreme diversity of TSF subgroups relative to each other challenges our ability to make high quality alignments between subgroups. For example, the sequences of the founder 4-OT (30) and *cis*-CaaD (34) enzymes (Figure 2.1) are only 18% pairwise identical with each other. Second, the sparse experimental coverage of the superfamily severely limits our knowledge of the mechanistic and structural variations present in the TSF, or even the breadth of reaction space it supports. While these issues prevent a comprehensive study of TSF subgroup relationships, an initial examination of structure-function relationships between the best-studied members of the 4-OT and *cis*-CaaD subgroups was somewhat more tractable, as reported in this section.

#### 2.3.5.1 “Linkers” between *cis*-CaaD and 4-OT subgroup identify a similarity path between them

The SSN provided in Figure 2.2 shows two representative edges that link the 4-OT and *cis*-CaaD subgroups, indicating a sequence similarity bridge between them. We identified representative nodes that formed the shortest path between the two founder sequences in each subgroup, highlighted in the left panel of Figure 2.5. This path is anchored at either end by the representative nodes in which the founder 4-OT and *cis*-CaaD enzymes are located, with the linker nodes along proceeding from founder 4-OT to Fused 4-OT, to Linker 2, to Linker 1, to CgX to founder *cis*-CaaD. (CgX is a structurally characterized *cis*-CaaD homolog with unknown function (33). It is an inefficient *cis*-CaaD as it has much lower catalytic efficiency for this reaction than does *cis*-CaaD and it does not exhibit absolute specificity for the *cis*-isomer.)

As small numbers of edges linking highly diverse subgroups in functionally diverse superfamilies may be statistically suspect (35, 38), the links between the representative nodes

shown in Figure 2.5 were examined further using more detailed SSNs. The results indicated that 48 statistically significant edges indeed link the 4-OT and *cis*-CaaD subgroups (see section 2.5 Experimental Procedures: *Creation of linker control networks* for details).



**Figure 2.5 Sequences in the 50% representative network that link the 4-OT and *cis*-CaaD subgroups.** Left. Blow-up of the region in the 50% representative network (see Figure 2.2) that links the 4-OT and *cis*-CaaD subgroups. Representative nodes are colored as in Figure 2.2 except that those containing founder 4-OT (labeled 4-OT) and *cis*-CaaD (labeled CC) sequences are enlarged and colored dark blue and red, respectively. The nodes containing linker sequences Fused 4-OT (labeled f4-OT), Linker 2 (labeled 2), Linker 1 (labeled 1), and CgX are also enlarged. Right. Structure-guided multiple sequence alignment of proteins labeled in the SSN blow-up on the left. Structures used: 4-OT, 1BJP; Fused 4-OT, 6BLM; Linker 2, 5UNQ; Linker 1, 5UIF; CgX, 3N4G; *cis*-CaaD, 2FLZ. Catalytic residues of the founder 4-OT, Pro-1, Arg-11 and Arg-39, (positions 69, 79, and 108 according to the numbering of the MSA, respectively) are boxed in blue, and catalytic residues of *cis*-CaaD, Pro-1, His-28, Arg-70, Arg-73, Tyr-103 and Glu-114, are boxed in red (positions 1, 28, 76, 79, 110, 121, according to the numbering of the MSA, respectively). Note that the short founder 4-OT aligns best with the second  $\beta$ - $\alpha$ - $\beta$  domain of the five other proteins in the figure. The alignment shows that the active site composition of the linker proteins becomes more *cis*-CaaD-like across the similarity path shown by the network. (Some catalytic residues in 4-OT and *cis*-CaaD come from different subunits, as described in the legend for Figure 2.6.)

A structure of a protein from each of these representative linker nodes was available, enabling creation of a high confidence structure-guided MSA that could provide insight about the structural variations along this path. The right panel of Figure 2.5 shows the structure-guided sequence alignment computed for these structures. The percent identity for each pair of sequences along the similarity path (right panel) is 42% between 4-OT and the Fused 4-OT, 41% between Fused 4-OT and Linker 2, 26% between Linker 2 and Linker 1, 31% between Linker 1

and CgX, and 33% between CgX and *cis*-CaaD. These similarities are much higher than the 18% sequence identity obtained for a direct pairwise comparison between the founder 4-OT and *cis*-CaaD sequences in the alignment, which significantly enhances the quality of the alignment and the information it contains.

This MSA shows that the active site composition of the linker proteins becomes more *cis*-CaaD-like in a stepwise manner across the similarity path from the founder 4-OT to *cis*-CaaD. Starting from the short 4-OT, the transition to the Fused 4-OT reflects the fusion of the single  $\beta$ - $\alpha$ - $\beta$  unit represented by the short 4-OTs to form proteins with two  $\beta$ - $\alpha$ - $\beta$  subdomains that are roughly twice as long. These short 4-OTs represent the majority of the proteins in the subgroup whereas the Fused 4-OTs are more like the majority of the proteins in the other TSF subgroups, nearly all of which are composed of two fused  $\beta$ - $\alpha$ - $\beta$  subdomains (Supplemental Figure 11). Among the characterized fusion proteins of the TSF, only three enzymes show substantial 4-OT activity (Fused 4-OT, Linker 2 (Table 2.1), and CHMI (56)). Supplemental Figure 6 indicates that the characterized Fused 4-OT protein matches best with the 4-OT Level 2 subgroup 1.

In the next step from Fused 4-OT, Linker 2 retains the 4-OT and Fused 4-OT-like sequence patterns, including conservation of Arg-39 (position 108 according to the numbering in the Figure 2.5 MSA), although it also exhibits some divergence away from them. Interestingly, as described in the section entitled “Additional linkers among other TSF subgroups,” both Linker 2 and the Fused 4-OT also show multiple structural similarities with other TSF subgroups. In contrast to Linker 2, Linker 1 shows significant *cis*-CaaD-like active site properties, including a conserved His-28, Arg-70, and Glu-114 (positions 28, 76, and 121 according to Figure 2.5 numbering), while losing conservation of 4-OT-like active site residues, in particular, Arg-39. Both CgX and *cis*-CaaD show completion of the remainder of the active site machinery required

for *cis*-CaaD activity, Tyr-103 and Glu-114 (positions 110 and 121 by Figure 2.5 numbering). Although they represent contemporary proteins, these linker sequences may provide hints about intermediate or “transitional” features resulting from divergence from ancestral genes. Based on the active site variations observed in the linker set, along with the enriched representation of archaeal sequences in the 4-OT subgroup, we speculate that the Fused 4-OT subgroup members, the linker set and founder *cis*-CaaD may have diverged from a short 4-OT-like ancestor.

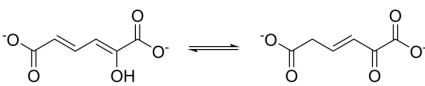
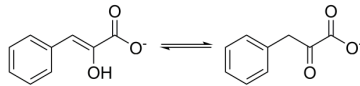
#### 2.3.5.2 Kinetic analysis shows loss of tautomerase activity in the enzymes linking 4-OT to *cis*-CaaD.

As a first step in investigating the functional consequences of the linker transitions, we examined the enzymes in the structure-guided MSA (Figure 2.5) for tautomerase activity using 2-hydroxymuconate (2-HM) and phenylolpyruvate (PP) (Table 2.1). These substrates were selected because both are processed by 4-OT, and PP is processed by *cis*-CaaD. (The tautomerase activity of *cis*-CaaD has been long regarded as an evolutionary vestige from a tautomerase-like ancestor (57).)

An examination of changes in active site architecture in the linker proteins (Figures 2.5 and 2.6), suggest that Linker 1, *cis*-CaaD, and CgX would not process a dicarboxylate substrate such as 2-HM. As a result, the monocarboxylate substrate PP would be expected to be more informative for gauging changes in tautomerase activity of the linker proteins. Indeed, both Fused 4-OT and Linker 2 are proficient tautomerases with 2-HM, whereas Linker 1, *cis*-CaaD, and CgX show no activity. Compared to 4-OT, Linker 2 exhibits only a 2-fold drop in  $k_{cat}/K_m$ , while Fused 4-OT shows a 23-fold drop in  $k_{cat}/K_m$ . Nonetheless, both are highly efficient tautomerases using 2-HM. In contrast, tautomerase activity with PP is detected for all of the enzymes. There is a stepwise loss of this activity going from 4-OT and the 4-OT-like enzymes

(Fused 4-OT and Linker 2), to Linker 1, CgX and ultimately *cis*-CaaD. Especially noteworthy,

**Table 2.1 Tautomerase activity across the linker proteins using 2-hydroxymuconate (left) and phenylolpyruvate (right) as substrates**

Enzyme						
	$k_{\text{cat}}$ (s <sup>-1</sup> )	$K_m$ (μM)	$k_{\text{cat}}/K_m$ (M <sup>-1</sup> s <sup>-1</sup> )	$k_{\text{cat}}$ (s <sup>-1</sup> )	$K_m$ (μM)	$k_{\text{cat}}/K_m$ (M <sup>-1</sup> s <sup>-1</sup> )
Founder 4-OT <sup>a</sup>	4000 ± 182	62 ± 8	(6.5 ± 0.9) × 10 <sup>7</sup>	73 ± 6	199 ± 23	(3.7 ± 0.5) × 10 <sup>5</sup>
Fused 4-OT	580 ± 25	208 ± 16	(2.8 ± 0.3) × 10 <sup>6</sup>	136 ± 8	720 ± 75	(1.9 ± 0.2) × 10 <sup>5</sup>
Linker 2	2800 ± 130	80 ± 9	(3.5 ± 0.4) × 10 <sup>7</sup>	920 ± 48	685 ± 68	(1.3 ± 0.2) × 10 <sup>6</sup>
Linker 1	N.D. <sup>b</sup>	N.D.	N.D.	80 ± 9	230 ± 40	(3.5 ± 0.7) × 10 <sup>5</sup>
CgX <sup>c</sup>	N.D.	N.D.	N.D.	7.5 ± 0.50	410 ± 40	(1.8 ± 0.2) × 10 <sup>4</sup>
Founder <i>cis</i> -CaaD	N.D.	N.D.	N.D.	0.20 ± 0.03 <sup>d</sup>	110 ± 30	(1.8 ± 0.6) × 10 <sup>3</sup>

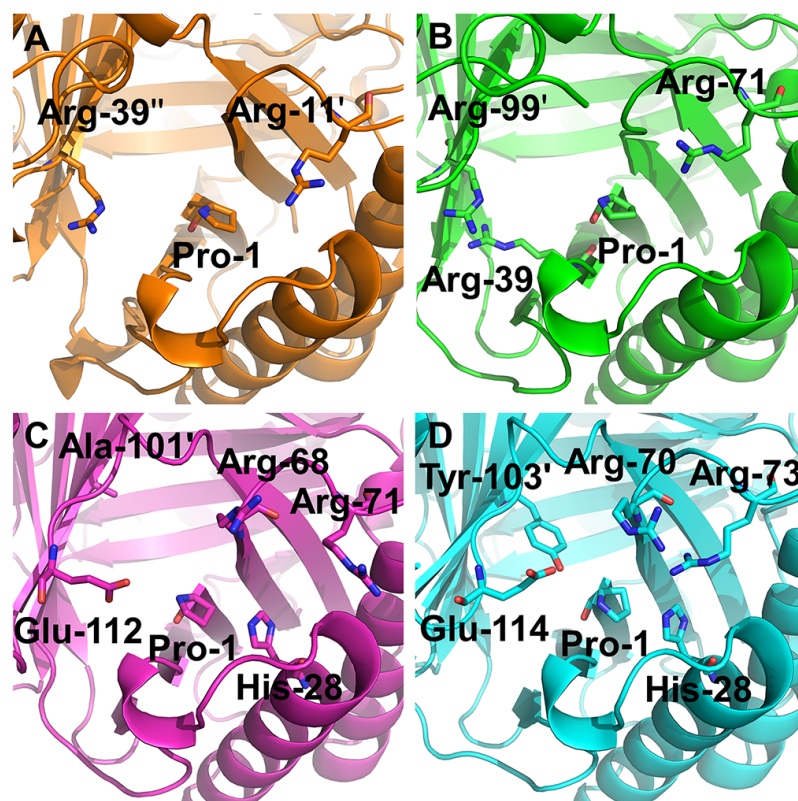
<sup>a</sup>(26).

<sup>b</sup>Not detected.

<sup>c</sup>CgX and *cis*-CaaD show *cis*-CaaD activity. Linker 2 shows a trace of *cis*-CaaD activity. Linker 1 and Fused 4-OT do not show *cis*-CaaD activity.

<sup>d</sup>(56).

the variations in specific functionally important residues correlate with this activity trajectory (Figure 2.6). Taking together the active site transitions illustrated in Figure 2.5, along with the kinetic data given in Table 2.1, these linker sequences may provide hints about ancestral features relevant to the evolution of the *cis*-CaaD and 4-OT-like proteins.



**Figure 2.6 Structural comparison of conserved active site residues in Linker 1 and Linker 2 with respect to the known catalytic residues of founder 4-OT and *cis*-CaaD.** The structures for each of these proteins are the same as were included in the structure-guided MSA shown in Figure 2.5. The unprimed, primed, and doubly primed residues indicate that they come from different subunits. A. Founder 4-OT. Pro-1 is positioned between Arg-11' and Arg-39''. This arrangement allows binding of the dicarboxylate substrate 2-HM by both arginine residues, and proton transfer by Pro-1 from the 2-hydroxyl group of 2-HM to C5. B. Linker 2. The active site architecture of Linker 2 is much like that of founder 4-OT. Arg-71 and Arg-99' (boxed in the MSA) are structurally equivalent to Arg-11' and Arg-39'' in founder 4-OT, respectively. Both arginine residues are present in the second  $\beta$ - $\alpha$ - $\beta$  subdomain of the Linker 2 monomer, which explains their very different position in the protein sequence. Linker 2 also has an Arg-39 in its first  $\beta$ - $\alpha$ - $\beta$  subdomain, which structurally forms part of the wall of the active site near Arg-99'. Its proximity to both Arg-99' and Pro-1, could signify a potential role in catalysis. C. Linker 1. Linker 1 exhibits an active site architecture very different from that of the founder 4-OT and Linker 2, and instead is more similar to the active site of *cis*-CaaD (panel D). One important difference is the absence of an arginine residue that is structurally equivalent to Arg-39'' in founder 4-OT and Linker 2. Instead, this arginine residue (Arg-68) appears to be repositioned much closer to Arg-71 (structurally equivalent to Arg-11' and Arg-71 in founder 4-OT and Linker 2, respectively). The other residues highlighted, His-28, Ala-101', and Glu-112, are structurally equivalent to His-28, Tyr-103', Glu-114 in founder *cis*-CaaD. Except for missing Tyr-103', which has Ala-101' in that position, the catalytic machinery of founder *cis*-CaaD is complete in Linker 1. D. Founder *cis*-CaaD. The active site of *cis*-CaaD shows its known catalytic machinery, composed of Pro-1, His-28, Arg-70, Arg-73, Tyr-103' and Glu-114.

### 2.3.6. Phylogenetic reconstruction of the linker set

To examine the relationships across the linker set using an independent approach, we generated a phylogenetic tree. Guided by SSNs created to expand the linker set to include specific proteins from the representative nodes shown in Figure 2.5 as well as their closest homologs, we identified 63 sequences for inclusion in the tree (see section 2.5 Experimental Procedures). The MSA comprised of these sequences that was used to compute this tree is available as Supplemental Figure 14. The unrooted phylogenetic tree of these proteins is provided in Figure 2.7. (Short 4-OT sequences were not included in the tree because these sequences contain only about half as much information as the rest of the fused sequences of the linker set, raising complications for interpreting the tree.)

The probability values at the leaf branches of the tree are generally of high quality within most of the major clades, although the diversity of the sequences is too great to generate an MSA of sufficiently high confidence to support rooting the tree. However, even the poorer significance values at the interior nodes provide support for the relationships along the similarity path given in Figure 2.5. That data, along with the tree and the kinetic results (Table 2.1), are generally consistent with each other, providing added confidence in our approach for dissecting details of the structure-function relationships between the 4-OT and *cis*-CaaD subgroup enzymes. Without the global context provided by a TSF-wide SSN, identification and analysis of the linker set would have been difficult to achieve.





the likely (by parsimony) emergence of Fused 4-OTs from a simpler ancestral type composed of a single  $\beta$ - $\alpha$ - $\beta$  subdomain and the enrichment of archaeal sequences in the short 4-OT subgroup relative to the lack of archaeal members in the *cis*-CaaD subgroup. We cannot, however, assume a unidirectional evolution from 4-OT to *cis*-CaaD from that data alone. Even though the phylogenetic tree provides new insight into the relationships between these two subgroups, the tree cannot be used to resolve the direction of their evolution, both because the tree is unrooted and because the short 4-OT proteins are not included in the tree.

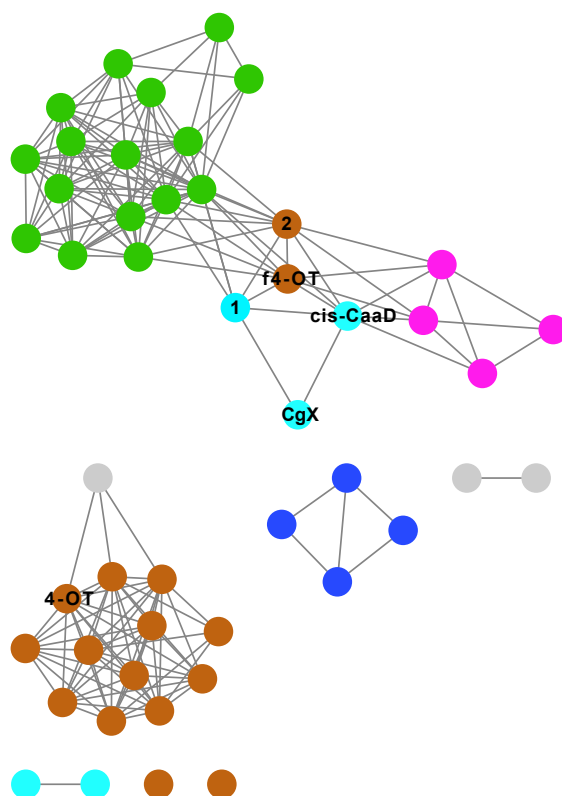
Interestingly, the phylogenetic tree suggests additional groups of sequences that might be added to the linker set originally identified in this work. In the absence of available structures for proteins in these diverse unlabeled clades of the tree, identification of active site variations that might offer new clues about the functions of these proteins remains beyond the scope of this work. The positions of these sequences in the tree do, however, suggest targets for structural characterization that might help to fill out the similarity path in more detail. As a case in point, the identification of Linker 1 led to our targeting of that protein for structural characterization (PDB 5UIF) (58), allowing us to include it in the structure-guided alignment shown in Figure 2.5

### **2.3.7. Additional linkers among other TSF subgroups**

Further studies to gain insight about structure-function relationships across the TSF would undoubtedly benefit from studying a larger context that includes all of the subgroups. For the work reported here, the extreme sequence diversity among the five TSF subgroups along with the scarcity of structures across the superfamily prevented creation of a high confidence MSA sufficient to support phylogenetic reconstruction of the entire superfamily. Characterization of additional structures that better sample all of the TSF subgroups would aid in generating a TSF-wide MSA. In the future, the addition of more TSF sequences from

metagenomic data may also contribute to creation of a statistically significant superfamily-wide tree.

In addition to linkers between the 4-OT and *cis*-CaaD subgroups, the representative SSN shown in Figure 2.2 hints at additional putative links among other TSF subgroups, although the significance of these few linking edges is uncertain without further analysis such as reported here for the 4-OT and *cis*-CaaD linker set. As a next step toward expanding this larger context beyond the 4-OT and *cis*-CaaD subgroups, we used the currently available structures to compute a structure similarity network across the TSF. As structure similarity can be identified at greater levels of sequence diversity than can be achieved from sequence comparisons (59). Based on this network, an initial, albeit cursory view of those subgroup relationships is shown in Figure 2.8.



**Figure 2.8 Structure similarity network of the TSF.** Each node represents a single structure, colored by subgroup as in Figure 2.2. The threshold for drawing edges between structures is a TM-Align score of 0.8. (A TM-align score of 0.5 is suggested to be statistically significant (60).) This network contains 48 structures for which the PDB

accession codes are listed in A.16. Labeled nodes refer to proteins identified in the 4-OT/*cis*-CaaD linker set shown in Figure 2.5.

This network reveals connections between subgroups that are not readily apparent from the TSF SSN (Figure 2.2). Especially interesting, two of the key structural connections linking the 4-OT and *cis*-CaaD subgroups, Fused 4-OT and Linker 2, also exhibit multiple links to the MIF and MSAD subgroups, as well as links between the *cis*-CaaD subgroup and MIF and MSAD subgroups. Thus, the structure similarity network extends the context by which we can hypothesize linkers among most of the subgroups, with the CHMI subgroup currently representing an outlier subgroup, (as was also found in the SSN (Figure 2.2)). We speculate that the central position of Fused 4-OT and Linker 2 in the structure similarity network offers a starting point for new studies aimed at addressing the evolution of the contemporary TSF subgroups from an ancient scaffold.

The 4-OT composed of short single  $\beta$ - $\alpha$ - $\beta$  domain proteins do not connect with the Fused 4-OT structure node or with linker 2 in Figure 2.8. As with the phylogenetic tree, the lower information content of the 4-OTs complicates interpretation of their relationships to other linkers and subgroups in the structure similarity network and likely results in its separation in the layout shown in the figure. Thus, we cannot easily interpret the relationship between the short 4-OTs and the Fused 4-OTs with other linker sequences from either the sequence or structural similarity network. This issue is also reflected in and addressed in different ways in other analyses presented in this work.

## 2.4 Summary

The results of this analysis show, for the first time, a global view of structure-function relationships across the TSF. As with other superfamilies that have been computationally

characterized in recent years, a large majority of the TSF members have not been experimentally characterized in any way, so that generating high confidence functional hypotheses for unknowns based on similarity to the few available “knowns” remains difficult and likely to be plagued by high levels of misannotation (61). Surprisingly, our global analysis of known TSF sequences reveals that about 3% of the superfamily sequences lack an N-terminal proline, raising important questions about the long-held assumption that a Pro-1 residue is required for catalytic activity. Guided by structure-based alignments, we identify “linker” sequences that best connect active site variations between the 4-OT and *cis*-CaaD subgroups, providing an indication of structural “transitions” that may distinguish the different reaction classes in each. The type of linker analysis described in this work offers a more information-rich strategy to compare variations between contemporary homologous proteins, providing more detailed information and contributing to biological and mechanistic insight in ways that are not so accessible from simple pairwise comparisons. Phylogenetic reconstruction of these linker relationships is consistent with the network-based linker analyses. Finally, structure-based identification of links among most of the TSF subgroups defined in this work provides new directions for examining how each subgroup may have evolved in different ways to produce new catalytic and biological functions.

## **2.5 Experimental Procedures**

### *2.5.1.1 Sequence Data Collection*

The tautomerase superfamily (TSF) has sequence entries annotated as tautomerasases in many of the major protein databases. InterPro (62) identifies the TSF by the following signatures: IPR028116, IPR001398, IPR015017, IPR004220, and IPR014347. These sequences were added to the Structure-Function Linkage Database (SFLD) (2) on August 16, 2016.

Following removal of redundant sequences, a final, non-redundant set of 11,395 sequences was used for further analyses.

#### *2.5.1.2 Construction of Sequence Similarity Networks*

Sequence similarity networks were constructed for the TSF according to algorithms from the Pythoscape software (36), tailored for use with the SFLD and its hardware environment. Two types of SSNs were computed: a one-sequence-per-node network (also called a 100% identity network) and representative SSNs. For the former, the sequences were pairwise compared in an all-by-all manner using the BLAST algorithm (63) with an edge recorded if BLAST reported an alignment with an *E*-value more significant than  $10^{-5}$ . For the figures provided in this work, edges were further pruned by thresholding the *E*-value score at which edges were drawn. The representative networks were built using the same BLAST data, but each node represents one or more TSF sequences identified by the CD-Hit algorithm (64) at a pairwise percent identity value as reported for each SSN figure. This reduction of many similar nodes into single representative nodes allowed creation of representative networks that could be analyzed at several levels of detail based on the percent identity cutoff chosen for binning sequences in each representative node. For this study, large networks were generated for the TSF using percent identity cutoffs at  $\geq 50\%$ ; for smaller subgroups or subsets of TSF sequences SSNs were computed using  $\geq 90\%$ , or 100% pairwise identity cutoffs. For each percent identity cutoff, the number of sequences in the representative nodes can be highly variable, as reported in the text. A representative edge between each pair of representative nodes was recorded if the one-sequence-per-node network recorded any edges between sequences in one representative node with sequences in the other. The score for each representative edge was computed as the geometric mean of the BLAST *E*-

value scores of all recorded one-sequence-per-node edges between the connected representative nodes.

Networks were analyzed using a thresholded approach such that edges were drawn between each pair of nodes if the similarity between those sequences was better than a statistical significance threshold chosen to illustrate the similarity relationships in the SSN at different levels of detail. The edge metric used for these comparisons was the BLAST  $E$ -value (used as a score (35)). This thresholded approach allowed examination of SSNs across a range of  $E$ -values. For the network(s) shown in the 50% representative network depicted in Figure 2.2 (and in Figures 2.3, 2.4A, and 2.5) the set of thresholds ( $E$ -value scores) was manually examined (between  $10^{-2}$  to  $10^{-20}$ ) to identify a threshold estimated to be near optimal for displaying the relationships in each SSN. All of the networks shown in this work were visualized in Cytoscape (65), using the Organic layout. The length of edges using this layout correlates with order of connectivity, and generally track with dissimilarity (35). Note: The representative nodes comprising the MIF subgroup in the SSN figures presented Figures 2.2, 2.3, 2.4A, and 2.5 were moved slightly because they were overlaid on an edge connecting the MSAD and 4-OT subgroups, giving the erroneous inference that that edge also connected to the MIF subgroup. As edges visualized with the Organic layout do not directly denote similarity distance but instead correlate with the number of pairwise connections between nodes, this manipulation does not alter the quantitative representation of the data.

#### *2.5.1.3 Division of the TSF into subgroups based on sequence similarity*

An optimal threshold of  $10^{-11}$  was chosen for subgrouping the sequences of the TSF into Level 1 subgroups in which the similarities in each were greater within each subgroup than between subgroups. The degree to which each subgroup is distinct from other subgroups (at the

*E*-value threshold at which it was visualized) was verified using MSAs and HMMs (37) as described in the text. Sequence clusters that were too small or too diverse for generation of a useful MSA and HMM were not further named or classified.

As the diversity of the Level 1 4-OT subgroup was too great to allow creation of a single high quality MSA depicting the entire subgroup, these sequences were further subgrouped guided by a new and more detailed representative network generated from the Level 1 4-OT subgroup. Each representative node was comprised of the 4-OT sequences binned at  $\geq 90\%$  pairwise identical. This network is visualized at an *E*-value threshold of  $10^{-18}$  and shown in Supplemental Figure 6. Using a procedure similar to that described for generation of the Level 1 subgroups, four of the largest clusters were defined as Level 2 subgroups (denoted in the SFLD as Groups 1-4 and listed below the Level 1 4-OT subgroup). MSAs and HMMs for these Level 2 subgroups verified their distinctness using a similar procedure to that used to validate the Level 1 subgroups.

#### *2.5.1.4 Construction of the Structure Similarity Network*

In analogy to sequence similarity networks, a structural similarity network for the TSF was created using the TM-Align algorithm (66) to perform pairwise comparisons among a non-redundant set of TSF structures (Supplemental File 1). The crystal structures used for input to TM-Align were identified by searching the PDB using Pfam (67) signatures PF14832, PF01187, PF08921, PF02962, PF01361, and PF14552. The resulting networks were then visualized in Cytoscape. A cutoff score of 0.5 or higher is considered to be statistically significant by the authors of the TM-Align algorithm (60). A cutoff score  $\geq 0.8$  was used for Figure 2.8, well above the 0.5 score. To aid in comparing the topologies of the SSN and the structure similarity



network, the 0.8 cutoff was chosen to match to the extent possible the subgroup boundaries used in visualizing the Level 1 SSN.

#### 2.5.1.5 Creation of HMMs

An MSA was generated for each named Level 1 and Level 2 subgroup using Clustal Omega (68) and manually refined. HMMs were created using HMMER 3 (69). These subgroup-specific HMMs were mapped to the relevant representative networks to enable visual estimation of how well each captured the representative nodes of the subgroup and the degree to which each overlapped with other subgroups.

#### 2.5.1.6 Identification of TSF members without an N-terminal Proline

Computational analysis of the 11,395 non-redundant sequences for those without an N-terminal Met-Pro yielded 2,296 sequences. The 495 eukaryotic sequences were removed leaving 1801 prokaryotic sequences. Manual examination of these sequences for a correctly annotated start codon in frame with a ribosome binding site (RBS) resulted in the set of 346 sequences without an N-terminal proline. An explanation and examples of the manual curation process are provided in Supplemental Figure 15.

#### 2.5.1.7 Identification of linkers

Two edges were observed in the 50% representative network (Figure 2.2) that connect the founder 4-OT and *cis*-CaaD subgroups. These two linking edges are defined by an almost identical mean BLAST *E*-value of  $10^{-11.015}$  and  $10^{-11.111}$ , respectively. One sequence from each representative node connected by the edge defined by a geometric mean BLAST *E*-value of  $10^{-11.015}$  was selected for experimental characterization. The representative node in the founder 4-OT subgroup contains only one sequence (Linker 2: UniProt F4GMX9). The representative node

in the *cis*-CaaD subgroup contains three sequences. The best representative of these sequences (Linker 1: UniProt K9NIA5) was identified by creating an MSA and HMM for that subgroup, and then selecting the sequence that scored highest against that HMM. Both proteins, designated respectively Linker 1 and Linker 2, were structurally characterized (see *Structure characterization* section below). The representative nodes for the second edge connecting the 4-OT and *cis*-CaaD subgroups from the 50% network were named N1 and N2. N1 and N2 were included in the linker control experiment (Supplemental Figure 13) and N1 was included in the phylogenetic tree, as described below. Within the 4-OT and *cis*-CaaD subgroups, the nodes containing Fused 4-OT (PDB: 6BLM) and CgX (a *cis*-CaaD homologue from *Corynebacterium glutamicum* of unknown function (33)) (PDB: 3N4G) were identified as linkers that connect the representative nodes of Linker 2 and Linker 1 to those that contain the founder 4-OT (PDB: 1BJP) and *cis*-CaaD (PDB: 2FLZ) proteins, respectively.

#### 2.5.1.8 Creation of linker control networks

To determine the number of individual edges included in the representative nodes of the linker set, we constructed a 90% representative network (not shown) including all members of the *cis*-CaaD and 4-OT subgroups represented in Figure 2.2. Twenty-five edges were observed that link 16 representative nodes in the founder 4-OT subgroup to 13 representative nodes in the *cis*-CaaD subgroup (generating a total of 29 representative linker nodes between the two subgroups at this threshold). The geometric mean BLAST *E*-value of these 25 representative edges ranged from  $10^{-11.015}$  to  $10^{-13.046}$ . The first neighbor nodes (one hop out) of these 29 representative nodes, along with their first neighbor representative nodes of these representative nodes were selected (in total, two hops out from the 29 linker representative nodes). A one sequence per node network was then generated from the selected representative linker nodes

composed of 2,761 non-redundant sequences, as shown in Supplemental Figure 13. Forty-eight edges were observed that link 31 nodes from the founder 4-OT subgroup to 17 nodes in the *cis*-CaaD subgroup. The BLAST *E*-value of these 48 edges ranged from  $10^{-11.046}$  to  $10^{-13.046}$ , indicating that multiple statistically significant edges link the 4-OT and *cis*-CaaD subgroups.

#### 2.5.1.9 Structure characterization

The details for the crystallization and structure determination of Fused 4-OT and Linker 2 (inactivated by the irreversible inhibitor, 2-oxo-3-pentynoate (29), will be reported in a future publication. Briefly, crystals were obtained by the sitting drop vapor-diffusion method at room temperature. X-ray diffraction data were collected at the Advanced Light Source beamline 5.0.3 (ALS, Berkeley, CA) with a wavelength of 0.97741 Å (Fused 4-OT) or 0.97641 Å (Linker 2) at 100 K. Structures were determined using molecular replacement where the Linker 2 (inactivated by 2-oxo-3-pentynoate) structure (PDB entry 5UNQ, 40% sequence identity) was used for Fused 4-OT and the 4-OT structure (PDB entry 1BJP, 37% and 38% sequence identity with the N-terminal and C-terminal sequence, respectively) was used for Linker 2 inactivated by 2-oxo-3-pentynoate. The Linker 2 structure showed modified and unmodified active sites in different chains (of three chains per biological unit). The unmodified chain of Linker 2 was used in guiding the construction of the MSA shown in Figure 2.5 as well as in the structure similarity network shown in Figure 2.8. Both structures were deposited in the PDB: Fused 4-OT, 6BLM; Linker 2, 5UNQ.

#### 2.5.1.10 Creation of the structure-guided alignment

Chimera (70) was used to make a structure guided MSA for the founder 4-OT and *cis*-CaaD sequences with the Fused 4-OT, CgX, and linker protein sequences. Structures used were founder 4-OT (PDB: 1BJP), Fused 4-OT (6BLM), Linker 2 (PDB: 5UNQ), Linker 1 (PDB:

5UIF), CgX (PDB: 3N4G), and *cis*-CaaD (PDB: 2FLZ). Chimera's MatchMaker tool was used to align the structures, using CgX (PDB: 3N4G) as the reference structure

#### 2.5.1.11 Phylogenetic tree construction of the 4-OT and *cis*-CaaD subgroups

As the time requirement for computation of phylogenetic trees can become unrealistic for highly diverse sequence sets, the 90% representative network made up of 3,006 representative nodes was pruned to provide a more realistic number of sequences for generating the tree. This was achieved by first generating a 70% identity representative network composed of 1,496 representative nodes that included all the sequences from the 4-OT and *cis*-CaaD subgroups. First neighbor nodes (one hop out) of the 70% representative nodes containing Linker 1, Linker 2, linkers N1 and N2 were selected resulting in a total of 121 representative nodes. Subsequently, the sequence with the highest score to the associated HMM created for each respective representative node was selected for inclusion in the tree. From this set of 121 sequences, the 63 sequences composed of fused  $\beta$ - $\alpha$ - $\beta$ -subdomains (sequence length >110) were selected to be used in the phylogenetic tree.

A structure-guided MSA of these 63 sequences was generated using Chimera. Chimera's MatchMaker tool was used to align the structures, using *cis*-CaaD (PDB: 2FLZ) as the reference structure. To create the MSA, the shortest path on the 90% representative network was calculated between the linker nodes of the 4-OT and *cis*-CaaD subgroups and the sequences representing Linker 1, Linker 2 or Fused 4-OT. That is, the subset of those 63 sequences that belonged to the *cis*-CaaD subgroup in the 90% representative network, were aligned to a superposition of the structures of CgX and *cis*-CaaD. Similarly, the remainder of these 63 sequences, which belonged to the 4-OT subgroup, was aligned to a superposition of the structures of Fused 4-OT, and Linker

2. Once this MSA was constructed, it was realigned using MUSCLE (71) to better resolve gapped regions and alignment of variable regions among the sequences near the C-termini.

This structure-guided MSA (Supplementary Figure S14) was used as input for computing the tree, which was calculated using the Mr Bayes software (72). The tree was calculated using one Metropolis-coupled mcmc chains from 2 runs using 1,000,000 generations, sampled every 100 generations. The tree was visualized via FigTree using the radial layout (available from <http://tree.bio.ed.ac.uk/software/figtree/>).

#### 2.5.1.12 Kinetic analysis of linkers

The substrates were prepared and kinetic parameters were determined as described elsewhere (30). Nonlinear regression data analysis was performed using the program Grafit (Erithacus Software Ltd., Staines, U.K.). Linker 1, Linker 2, and Fused 4-OT were produced by tailoring a previous protocol as needed to obtain sufficient amounts of soluble protein at the purity required for kinetic measurements (30). The genes encoding Linker 1 from *Pseudomonas* sp. UW4 (UniProt K9NIA5), Linker 2 from *Pusillimonas* sp. (strain T7-7) (UniProt F4GMX9) and Fused 4-OT from *Burkholderia lata* (strain ATCC 17760) (UniProt Q392K7) were codon-optimized for their separate expression in *E. coli*, synthesized and cloned into the expression vector pJ411 by DNA2.0, Inc., now A TUM (Newark, CA).

#### Acknowledgments

The work in this paper was supported by the Robert A. Welch Foundation (F-1334 to CPW and F-1778 to YJZ) and the National Institutes of Health Grants GM-41239 (CPW) and GM-60595 (PCB). Instrumentation and technical assistance were provided by the Macromolecular Crystallography Facility, with financial support from the College of Natural Sciences, the Office of the Executive Vice President and Provost, and the Institute for Cellular and Molecular

Biology at the University of Texas at Austin. The X-ray diffraction data were collected at the Advanced Light Sources (ALS) (Berkeley, CA). The Berkeley Center for Structural Biology is supported in part by the National Institutes of Health, National Institute of General Medical Sciences, and the Howard Hughes Medical Institute. The Advanced Light Source is supported by the Director, Office of Science, Office of Basic Energy Sciences, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

The authors are indebted to Marieke Baas for carefully proofreading and formatting the manuscript and to Kathy Clement for help with generating graphics files for SSNs.

Database names and accession codes: The atomic coordinates and structure factors have been deposited as Protein Data Base entries 5UIF (Linker 1), 5UNQ (Linker 2), and Fused 4-OT (6BLM).

### **Author Contributions**

PCB conceived and coordinated the computational work and CPW conceived and coordinated the experimental work. Both provided data analysis in their respective areas and both wrote the paper. RD conceived computational approaches for analysis and comparison of superfamily data, developed and performed bioinformatics analyses, curated the superfamily in the SFLD and wrote parts of the paper. BJB generated the enzymes, and carried out the kinetic analysis of Fused 4-OT, Linker 1, and Linker 2. BJB also assisted in the computational and phylogenetic analysis and wrote parts of the paper. JAL and YJZ coordinated the crystallographic work and determined the X-ray structures of Fused 4-OT and Linker 2. CRP assisted in the determination of the kinetic parameters in Table 1. EA provided expertise and guidance for phylogenetic inference and performed the analysis of the MIF one sequence per node subgroup. BJP updated the SFLD backend as required for this project, generated networks for download from the SFLD,

and contributed to the design of the linker control experiment. GLH generated an initial set of TSF representative sequences, added them to the SFLD, and aided in creation of subgroup HMMs.

## Chapter 3. Curation of the Tautomerase Superfamily

### 3.1 Introduction

A Superfamily is a set of evolutionarily related proteins with a shared chemical capability that maps to a set of conserved active site features. They can be highly divergent and catalyze many overall reactions with different mechanisms. Such superfamilies are termed functionally or mechanistically diverse and tend to exhibit complicated structure-function relationships. In the Structure Function Linkage Database (SFLD), developed by the Babbitt lab, these superfamilies are the focus of the Core Superfamily set . The SFLD describes superfamilies in a hierarchical manner. The family is the finest level of detail, at which all proteins are often closely related and all perform the same overall chemistry using the same mechanistic strategy. However, sometimes proteins in the same family are very diverse, likely not participating in the same overall chemistry or mechanistic strategy. In the subgroup (sometimes referred to as a Level-1 subgroup), which is the least granular level of detail, we place sequences that are more similar to one another than they are to sequences in another subgroup. Here, Level-1 subgroups of the TSF are presented as well as Level-2 subgrouping of a Level-1 subgroup.



SFLD - Superfamily - Tautome x

sflid.rbvi.ucsf.edu/django/superfamily/159/

UCSF University of California, San Francisco | About UCSF | UCSF Medical Center

SFLD RBVI

Home About SFLD Documentation Tutorials Contact Us Curator's Entrance

Browse by Superfamily Browse by Reaction Search by Enzyme Search by Reaction

**Top Level** Name  
Superfamily (core) Tautomerase

	Total	Family known		Family unknown
		100%	<100%	
Functional domains	14945	0	0	14945
UniProtKB	29992	0	0	29992
GI	49171	0	0	49171
Structures	148			
Reactions	0			

Functional domains of this superfamily were last updated on Nov. 19, 2017  
New functional domains were last added to this superfamily on Aug. 17, 2016

Description References (1) Curator Notes

All biochemically characterized enzymes in the superfamily fall into five reaction types (see Figure 1 of Davidson et. al.), each of which uses the N-terminal proline either as a general base or a general acid. All previous studies of characterized TSF members show a shared utilization of an N-terminal proline in their mechanisms, leading to an expectation that all TSF members would exhibit this feature. Notably, the common reaction catalyzed by three of these enzymes, 4-oxalocrotonate tautomerase (4-OT), 5-(carboxymethyl)-2-hydroxymuconate isomerase (CHMI), and the phenylpyruvate tautomerase (PPT) activity of macrophage migration inhibitory factor (MIF) is an enol-keto tautomerization of a pyruvoyl moiety in which the Pro-1 has a low pKa value (-6.4 in 4-OT). The two other reaction types, catalyzed by cis- and trans-3-chloroacrylic acid dehalogenase (cis-CaaD and CaaD, respectively) and malonate semialdehyde decarboxylase (MSAD), are more divergent in mechanism. While these latter two enzyme-catalyzed reactions still utilize the N-terminal proline, it has a higher pKa value (-9.2 in CaaD) and functions as a general acid. In addition to its PPT activity, MIF is more commonly recognized as functioning as a pro-inflammatory cytokine in mammals. The PPT and the thiol-protein oxidoreductase activities (21,22) of MIF are not involved in its cytokine activities.

Select Task → Download Network Download Data Set

Subgroup	T	K	C	U	S
4-oxalocrotonate tautomerase	4595	0	0	4595	26
└ 1: Group 1	1118	0	0	1118	14
└ 1: Group 2	366	0	0	366	0
└ 1: Group 3	1149	0	0	1149	2
└ 1: Group 4	332	0	0	332	3
5-(carboxymethyl)-2-hydroxymuconate isomerase	3908	0	0	3908	14
cis-3-chloroacrylic acid dehalogenase	159	0	0	159	0
macrophage migration inhibitory factor	3110	0	0	3110	97
malonate semialdehyde decarboxylase	2109	0	0	2109	9

A joint project supported by NSF-DBI-1356193 (P. Babbitt & G. L. Holliday, CoPIs), NIH R01GM60595 (P. Babbitt, PI), and NIGMS P41GM103311 (Resource for Biocomputing, Visualization, & Informatics) (T. Ferrin, PI). Additional support has been provided by NSF-DBI-0234768, NSF-DBI-0640476, NIGMS U54GM093342, and NIGMS P01GM071790.

Babbitt Lab, SFLD Team  
© 2004-2016 The Regents of the University of California. All rights reserved.

**Figure 3.1 The Tautomerase superfamily in the SFLD.** The Tautomerase Superfamily is presented as part of the core SFLD, comprised of 14,945 enzyme functional domains and 5 subgroups.

## 3.1 Level-1 Subgroups

### 3.1.1. 4-Oxalocrotonate Tautomerase

The screenshot shows the SFLD web interface for the 4-oxalocrotonate tautomerase subgroup. The page includes a navigation bar with links like 'Home', 'About SFLD', and 'Documentation'. Below the navigation, there are search options: 'Browse by Superfamily', 'Browse by Reaction', 'Search by Enzyme', and 'Search by Reaction'. The main content area displays the subgroup name and a summary table.

	Total	Family known <input type="checkbox"/>		Family unknown <input type="checkbox"/>
		100% <input type="checkbox"/>	<100% <input type="checkbox"/>	
Functional domains	4595	0	0	4595
UniProtKB	2434	0	0	2434
GI	3819	0	0	3819
Structures	26			
Reactions	0			

Functional domains of this subgroup were last updated on Nov. 19, 2017  
New functional domains were last added to this subgroup on Aug. 17, 2016

Description | References (1) | Curator Notes

The founding member of this group is experimentally characterized and found in a bacterial pathway for the degradation aromatic hydrocarbons. It exists as a hexamer of 62 amino acid subunits and converts 2-hydroxymuconate to 2-oxo-3-hexenedioate.

Select Task → | Download Network | Download Data Set

Subgroup Legend	T	K	C	U	S
4-oxalocrotonate tautomerase	4595	0	0	4595	26
└ 1: Group 1	1118	0	0	1118	14
└ 1: Group 2	366	0	0	366	0
└ 1: Group 3	1149	0	0	1149	2
└ 1: Group 4	332	0	0	332	3

A joint project supported by NSF-DBI-1356193 (P. Babbitt & G. L. Holliday, CoPIs), NIH R01GM60595 (P. Babbitt, PI), and NIGMS P41GM103311 (Resource for Biocomputing, Visualization, & Informatics) (T. Ferrin, PI). Additional support has been provided by NSF-DBI-0234768, NSF-DBI-0640476, NIGMS U54GM093342, and NIGMS P01GM071790.

Babbitt Lab, SFLD Team  
© 2004-2016 The Regents of the University of California. All rights reserved.

Figure 3.2 SFLD page for the 4-oxalocrotonate tautomerase subgroup.

### 3.1.2. 5-(carboxymethyl)-2-Hydroxymuconate Isomerase

UCSF University of California, San Francisco | About UCSF | UCSF Medical Center

SFLD RBVI

Home About SFLD Documentation Tutorials Contact Us Curator's Entrance

Browse by Superfamily Browse by Reaction Search by Enzyme Search by Reaction

Top Level Name

Superfamily (core) Tautomerase

Subgroup 5-(carboxymethyl)-2-hydroxymuconate isomerase

	Total	Family known		Family unknown
		100%	<100%	
Functional domains	3908	0	0	3908
UniProtKB	6787	0	0	6787
GI	11670	0	0	11670
Structures	14			
Reactions	0			

Functional domains of this subgroup were last updated on Oct. 27, 2017  
New functional domains were last added to this subgroup on Aug. 17, 2016

Description References (1) Curator Notes

The founding member for this group, UniProt Q05354, exists as a trimer that converts 5-carboxymethyl-2-hydroxy-muconic acid into 5-oxo-pent-3-ene-1,2,5-tricarboxylic acid. A CHMI monomer is approximately twice as long as a 4-OT monomer with a fold similar to that of a 4-OT dimer.

Select Task → Download Network View Alignment Align Sequence(s) Download Data Set

Subgroup Legend

Subgroup	T	K	C	U	S
5-(carboxymethyl)-2-hydroxymuconate isomerase	3908	0	0	3908	14

A joint project supported by NSF-DBI-1356193 (P. Babbitt & G. L. Holliday, CoPIs), NIH R01GM60595 (P. Babbitt, PI), and NIGMS P41GM103311 (Resource for Biocomputing, Visualization, & Informatics) (T. Ferrin, PI). Additional support has been provided by NSF-DBI-0234768, NSF-DBI-0640476, NIGMS U54GM093342, and NIGMS P01GM071790.

Babbitt Lab, SFLD Team  
© 2004-2016 The Regents of the University of California. All rights reserved.

Figure 3.3 SFLD page for the 5-(carboxymethyl)-2-hydroxymuconate isomerase subgroup.

### 3.1.3. *cis*-3-Chloroacrylic Acid Dehalogenase Subgroup

The screenshot shows the SFLD web interface for the *cis*-3-chloroacrylic acid dehalogenase subgroup. The page includes a navigation bar with links for Home, About SFLD, Documentation, Tutorials, and Contact Us. Below the navigation bar, there are search options: Browse by Superfamily, Browse by Reaction, Search by Enzyme, and Search by Reaction. The main content area displays a hierarchy: Top Level (Tautomerase) and Subgroup (*cis*-3-chloroacrylic acid dehalogenase). A table provides statistics for the subgroup, including Functional domains, UniProtKB, GI, Structures, and Reactions. The table also shows the number of family known and family unknown entries. A description of the enzyme is provided, along with a legend for the subgroup. The page footer contains funding information and contact details for the Babbitt Lab, SFLD Team.

	Total	Family known		Family unknown
		100%	<100%	
Functional domains	159	0	0	159
UniProtKB	172	0	0	172
GI	121	0	0	121
Structures	0			
Reactions	0			

Functional domains of this subgroup were last updated on Oct. 27, 2017  
 New functional domains were last added to this subgroup on Aug. 14, 2016

**Subgroup Legend**

	T	K	C	U	S
<i>cis</i> -3-chloroacrylic acid dehalogenase	159	0	0	159	0

A joint project supported by NSF-DBI-1356193 (P. Babbitt & G. L. Holliday, CoPIs), NIH R01GM60595 (P. Babbitt, PI), and NIGMS P41GM103311 (Resource for Biocomputing, Visualization, & Informatics) (T. Ferrin, PI). Additional support has been provided by NSF-DBI-0234768, NSF-DBI-0640476, NIGMS U54GM093342, and NIGMS P01GM071790.

Babbitt Lab, SFLD Team  
 sfld.rbvi.ucsf.edu/django/subgroup/1427/#tabs-tools-2 6 The Regents of the University of California. All rights reserved.

Figure 3.4 SFLD page for the *cis*-3-chloroacrylic acid dehalogenase subgroup

### 3.1.4. Macrophage Migration Inhibitory Factor Subgroup

The screenshot displays the SFLD web interface for the Macrophage Migration Inhibitory Factor subgroup. The page includes a navigation menu at the top with options like 'Home', 'About SFLD', 'Documentation', 'Tutorials', and 'Contact Us'. Below the navigation, there are search options: 'Browse by Superfamily', 'Browse by Reaction', 'Search by Enzyme', and 'Search by Reaction'. The main content area shows a breadcrumb trail: 'Top Level > Superfamily (core) > Subgroup'. A summary table provides statistics for the subgroup, including the total number of functional domains, UniProtKB entries, GI entries, structures, and reactions. A description of the Macrophage Migration Inhibitory Factor (MIF) enzymes is provided, along with a legend table showing the distribution of entries across various categories (T, K, C, U, S).

	Total	Family known		Family unknown
		100%	<100%	
Functional domains	3110	0	0	3110
UniProtKB	2082	0	0	2082
GI	3012	0	0	3012
Structures	97			
Reactions	0			

Functional domains of this subgroup were last updated on Oct. 27, 2017  
 New functional domains were last added to this subgroup on Aug. 17, 2016

**Subgroup Legend**

Subgroup	T	K	C	U	S
macrophage migration inhibitory factor	3110	0	0	3110	97

A joint project supported by NSF-DBI-1356193 (P. Babbitt & G. L. Holliday, CoPIs), NIH R01GM60595 (P. Babbitt, PI), and NIGMS P41GM103311 (Resource for Biocomputing, Visualization, & Informatics) (T. Ferrin, PI). Additional support has been provided by NSF-DBI-0234768, NSF-DBI-0640476, NIGMS U54GM093342, and NIGMS P01GM071790.

Babbitt Lab, SFLD Team  
 © 2004-2016 The Regents of the University of California. All rights reserved.

Figure 3.5 SFLD page for the macrophage migration inhibitory factor subgroup

### 3.1.5. Malonate Semialdehyde Decarboxylase Subgroup

UCSF University of California, San Francisco | About UCSF | UCSF Medical Center

SFLD RBVI Curator's Entrance

Browse by Superfamily Browse by Reaction Search by Enzyme Search by Reaction

Top Level Name  
 Superfamily (core) Tautomerase  
 Subgroup malonate semialdehyde decarboxylase

	Total	Family known		Family unknown
		100%	<100%	
Functional domains	2109	0	0	2109
UniProtKB	4315	0	0	4315
GI	7140	0	0	7140
Structures	9			
Reactions	0			

Functional domains of this subgroup were last updated on Oct. 27, 2017  
 New functional domains were last added to this subgroup on Aug. 17, 2016

Description References (1) Curator Notes

The founding member of this group, UniProt Q9EV83, converts malonate semialdehyde to acetaldehyde. It has also been implicated in the hydratase activity of MSAD in which 2-oxo-3-pentynoate is converted to acetopyruvate.

Select Task → Download Network View Alignment Align Sequence(s) Download Data Set

Subgroup Legend	T	K	C	U	S
malonate semialdehyde decarboxylase	2109	0	0	2109	9

A joint project supported by NSF-DBI-1356193 (P. Babbitt & G. L. Holliday, CoPIs), NIH R01GM60595 (P. Babbitt, PI), and NIGMS P41GM103311 (Resource for Biocomputing, Visualization, & Informatics) (T. Ferrin, PI). Additional support has been provided by NSF-DBI-0234768, NSF-DBI-0640476, NIGMS U54GM093342, and NIGMS P01GM071790.

Babbitt Lab, SFLD Team  
 © 2004-2016 The Regents of the University of California. All rights reserved.

Figure 3.6 SFLD page for the malonate semialdehyde decarboxylase subgroup.

# Level-2 Subgroups

## 3.1.6. 4-Oxalocrotonate Tautomerase – Group 1

The screenshot shows a web browser window with the URL `sfld.rbvi.ucsf.edu/django/subgroup/1428/`. The page header includes the UCSF logo and navigation links. The main content area displays a hierarchical tree structure for the subgroup, followed by a table of statistics and a description.

**Top Level**

- Superfamily (core): Tautomerase
- Subgroup: 4-oxalocrotonate tautomerase
- Group 1

	Total	Family known		Family unknown
		100%	<100%	
Functional domains	1118	0	0	1118
UniProtKB	8105	0	0	8105
GI	12717	0	0	12717
Structures	14			
Reactions	0			

Functional domains of this subgroup were last updated on Oct. 27, 2017  
New functional domains were last added to this subgroup on Aug. 17, 2016

**Description** | References (1) | Curator Notes

Several level-2 subgroups are defined due to the high amount of diversity among the sequences of the 4-OT subgroup (see supplemental figure 7 of Davidson et al.).

Select Task → | Download Network | View Alignment | Align Sequence(s) | Download Data Set

Subgroup	Legend	T	K	C	U	S
1: Group 1		1118	0	0	1118	14

A joint project supported by NSF-DBI-1356193 (P. Babbitt & G. L. Holliday, CoPIs), NIH R01GM60595 (P. Babbitt, PI), and NIGMS P41GM103311 (Resource for Biocomputing, Visualization, & Informatics) (T. Ferrin, PI). Additional support has been provided by NSF-DBI-0234768, NSF-DBI-0640476, NIGMS U54GM093342, and NIGMS P01GM071790.

Babbitt Lab, SFLD Team

Figure 3.7 SFLD page for Group 1 of the oxalocrotonate tautomerase subgroup

### 3.1.7. 4-Oxalocrotonate Tautomerase – Group 2

The screenshot shows the SFLD web interface for Group 2 of the oxalocrotonate tautomerase subgroup. The page includes a navigation menu, a hierarchical tree, a statistics table, a description, and a legend.

**UCSF** University of California, San Francisco | About UCSF | UCSF Medical Center

**SFLD** **RBVI**

Home About SFLD Documentation Tutorials Contact Us Curator's Entrance

Browse by Superfamily Browse by Reaction Search by Enzyme Search by Reaction

**Top Level** Name

- Superfamily (core) Tautomerase
- Subgroup 4-oxalocrotonate tautomerase
  - Group 2

	Total	Family known		Family unknown
		100%	<100%	
Functional domains	366	0	0	366
UniProtKB	680	0	0	680
GI	963	0	0	963
Structures	0			
Reactions	0			

Functional domains of this subgroup were last updated on Oct. 27, 2017  
New functional domains were last added to this subgroup on Aug. 17, 2016

Description References (1) Curator Notes

Several level-2 subgroups are defined due to the high amount of diversity among the sequences of the 4-OT subgroup (See supplemental figure 8 of Davidson et al.).

Select Task → Download Network View Alignment Align Sequence(s) Download Data Set

**Subgroup** Legend

	T	K	C	U	S
1: Group 2	366	0	0	366	0

A joint project supported by NSF-DBI-1356193 (P. Babbitt & G. L. Holliday, CoPIs), NIH R01GM60595 (P. Babbitt, PI), and NIGMS P41GM103311 (Resource for Biocomputing, Visualization, & Informatics) (T. Ferrin, PI). Additional support has been provided by NSF-DBI-0234768, NSF-DBI-0640476, NIGMS U54GM093342, and NIGMS P01GM071790.

Babbitt Lab, SFLD Team

Figure 3.8 SFLD page for Group 2 of the oxalocrotonate tautomerase subgroup



### 3.1.8. 4-Oxalocrotonate Tautomerase – Group 3

The screenshot shows the SFLD web interface for Group 3 of the 4-oxalocrotonate tautomerase subgroup. The page includes a navigation menu, a hierarchical tree, a statistics table, a description, and a legend.

**UCSF** University of California, San Francisco | About UCSF | UCSF Medical Center

**SFLD** **RBVI**

Home About SFLD Documentation Tutorials Contact Us Curator's Entrance

Browse by Superfamily Browse by Reaction Search by Enzyme Search by Reaction

**Top Level** Name

- Superfamily (core) Tautomerase
- Subgroup 4-oxalocrotonate tautomerase
  - Group 3

	Total	Family known		Family unknown
		100%	<100%	
Functional domains	1149	0	0	1149
UniProtKB	2111	0	0	2111
GI	3769	0	0	3769
Structures	2			
Reactions	0			

Functional domains of this subgroup were last updated on Oct. 27, 2017  
New functional domains were last added to this subgroup on Aug. 17, 2016

Description References (1) Curator Notes

Several level-2 subgroups are defined due to the high amount of diversity among the sequences of the 4-OT subgroup (See supplemental figure 7 of Davidson et al.).

Select Task → Download Network View Alignment Align Sequence(s) Download Data Set

**Subgroup** Legend

	T	K	C	U	S
1: Group 3	1149	0	0	1149	2

A joint project supported by NSF-DBI-1356193 (P. Babbitt & G. L. Holliday, CoPIs), NIH R01GM60595 (P. Babbitt, PI), and NIGMS P41GM103311 (Resource for Biocomputing, Visualization, & Informatics) (T. Ferrin, PI). Additional support has been provided by NSF-DBI-0234768, NSF-DBI-0640476, NIGMS U54GM093342, and NIGMS P01GM071790.

Babbitt Lab, SFLD Team

Figure 3.9 SFLD page for Group 4 of the oxalocrotonate tautomerase subgroup

### 3.1.9. 4-Oxalocrotonate Tautomerase – Group 4

The screenshot shows the SFLD web interface for Group 4 of the oxalocrotonate tautomerase subgroup. The page includes a navigation menu, a hierarchical tree, a statistics table, a description, and a legend.

**UCSF** University of California, San Francisco | About UCSF | UCSF Medical Center

**SFLD** **RBVI**

Home About SFLD Documentation Tutorials Contact Us Curator's Entrance

Browse by Superfamily Browse by Reaction Search by Enzyme Search by Reaction

**Top Level** Name

- Superfamily (core) Tautomerase
- Subgroup 4-oxalocrotonate tautomerase
  - Group 4

	Total	Family known		Family unknown
		100%	<100%	
Functional domains	332	0	0	332
UniProtKB	2181	0	0	2181
GI	3650	0	0	3650
Structures	3			
Reactions	0			

Functional domains of this subgroup were last updated on Nov. 19, 2017  
New functional domains were last added to this subgroup on Aug. 17, 2016

Description References (1) Curator Notes

Several level-2 subgroups are defined due to the high amount of diversity among the sequences of the 4-OT subgroup (See supplemental figure 9 of Davidson et al.).

Select Task → Download Network View Alignment Align Sequence(s) Download Data Set

Subgroup	Legend	T	K	C	U	S
1: Group 4		332	0	0	332	3

A joint project supported by NSF-DBI-1356193 (P. Babbitt & G. L. Holliday, CoPIs), NIH R01GM60595 (P. Babbitt, PI), and NIGMS P41GM103311 (Resource for Biocomputing, Visualization, & Informatics) (T. Ferrin, PI). Additional support has been provided by NSF-DBI-0234768, NSF-DBI-0640476, NIGMS U54GM093342, and NIGMS P01GM071790.

Babbitt Lab, SFLD Team

Figure 3.10 SFLD page for Group 4 of the oxalocrotonate tautomerase subgroup

## Chapter 4. Conclusion

As the amount of genomic data continues to grow and we in the scientific community continue to probe the functional relationships of mechanistically diverse enzyme superfamilies, the types of analysis shown in this work will become increasingly important. For example, the linker-guided strategy employed in this study can be used in the future analysis of other superfamilies to probe the sequence/structure/function transitions. For the first time a global view of the structure function relationships in the Tautomerase Superfamily is provided, revealing new insights and an initial set of hypothesis with some experimental validation. The scientific community can now build on these hypothesis in an iterative way, building additional insights about sequence/structure/function across the superfamily. The classification and interactive networks are readily available online on the Structure Function Linkage Database (<http://sfld.rbvi.ucsf.edu/django/superfamily/159/>). In this manner, other experimentalists in the community can easily access and iterate on the data. In this genomic age, this iterative network-based approach to generating functional hypothesis will become increasingly important to probe at and expand the limits of our knowledge about well-studied and new superfamilies.

## References

1. Davidson, R., Baas, B.-J., Akiva, E., Holliday, G. L., Polacco, B. J., LeVieux, J. A., Pullara, C. R., Zhang, Y. J., Whitman, C. P., and Babbitt, P. C. (2017) A global view of structure-function relationships in the tautomerase superfamily. *J. Biol. Chem.* 10.1074/jbc.M117.815340
2. Akiva, E., Brown, S., Almonacid, D. E., Barber, A. E., Custer, A. F., Hicks, M. A., Huang, C. C., Lauck, F., Mashiyama, S. T., Meng, E. C., Mischel, D., Morris, J. H., Ojha, S., Schnoes, A. M., Stryke, D., Yunes, J. M., Ferrin, T. E., Holliday, G. L., and Babbitt, P. C. (2014) The Structure-Function Linkage Database. *Nucleic Acids Res.* **42**, D521–30
3. Gerlt, J. A., and Babbitt, P. C. (2001) Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. *Annu. Rev. Biochem.* **70**, 209–246
4. Almonacid, D. E., and Babbitt, P. C. (2011) Toward mechanistic classification of enzyme functions. *Curr Opin Chem Biol.* **15**, 435–442
5. Murzin, A. G. (1996) Structural classification of proteins: new superfamilies. *Curr. Opin. Struct. Biol.* **6**, 386–394
6. Poelarends, G. J., Veetil, V. P., and Whitman, C. P. (2008) The chemical versatility of the beta-alpha-beta fold: catalytic promiscuity and divergent evolution in the tautomerase superfamily. *Cell. Mol. Life Sci.* **65**, 3606–3618
7. Stivers, J. T., Abeygunawardana, C., Mildvan, A. S., Hajipour, G., Whitman, C. P., and Chen, L. H. (1996) Catalytic role of the amino-terminal proline in 4-oxalocrotonate tautomerase: affinity labeling and heteronuclear NMR studies. *Biochemistry.* **35**, 803–813
8. Whitman, C. P. (2002) The 4-oxalocrotonate tautomerase family of enzymes: how nature makes new enzymes using a beta-alpha-beta structural motif. *Arch. Biochem. Biophys.* **402**, 1–13
9. Stivers, J. T., Abeygunawardana, C., Mildvan, A. S., Hajipour, G., and Whitman, C. P. (1996) 4-Oxalocrotonate tautomerase: pH dependence of catalysis and pKa values of active site residues. *Biochemistry.* **35**, 814–823
10. Subramanya, H. S., Roper, D. I., Dauter, Z., Dodson, E. J., Davies, G. J., Wilson, K. S., and Wigley, D. B. (1996) Enzymatic ketonization of 2-hydroxymuconate: specificity and mechanism investigated by the crystal structures of two isomerases. *Biochemistry.* **35**, 792–802
11. Lubetsky, J. B., Swope, M., Dealwis, C., Blake, P., and Lolis, E. (1999) Pro-1 of macrophage migration inhibitory factor functions as a catalytic base in the phenylpyruvate tautomerase activity. *Biochemistry.* **38**, 7346–7354
12. Stamps, S. L., Taylor, A. B., Wang, S. C., Hackert, M. L., and Whitman, C. P. (2000) Mechanism of the phenylpyruvate tautomerase activity of macrophage migration inhibitory factor: properties of the P1G, P1A, Y95F, and N97A mutants. *Biochemistry.* **39**, 9671–9678
13. Taylor, A. B., Johnson, W. H., Czerwinski, R. M., Li, H. S., Hackert, M. L., and

- Whitman, C. P. (1999) Crystal structure of macrophage migration inhibitory factor complexed with (E)-2-fluoro-p-hydroxycinnamate at 1.8 Å resolution: implications for enzymatic catalysis and inhibition. *Biochemistry*. **38**, 7444–7452
14. de Jong, R. M., Brugman, W., Poelarends, G. J., Whitman, C. P., and Dijkstra, B. W. (2004) The X-ray structure of trans-3-chloroacrylic acid dehalogenase reveals a novel hydration mechanism in the tautomerase superfamily. *J. Biol. Chem.* **279**, 11546–11552
15. Wang, S. C., Person, M. D., Johnson, W. H., and Whitman, C. P. (2003) Reactions of trans-3-chloroacrylic acid dehalogenase with acetylene substrates: consequences of and evidence for a hydration reaction. *Biochemistry*. **42**, 8762–8773
16. de Jong, R. M., Bazzacco, P., Poelarends, G. J., Johnson, W. H., Kim, Y. J., Burks, E. A., Serrano, H., Thunnissen, A.-M. W. H., Whitman, C. P., and Dijkstra, B. W. (2007) Crystal structures of native and inactivated cis-3-chloroacrylic acid dehalogenase. Structural basis for substrate specificity and inactivation by (R)-oxirane-2-carboxylate. *J. Biol. Chem.* **282**, 2440–2449
17. Poelarends, G. J., Serrano, H., Johnson, W. H., Hoffman, D. W., and Whitman, C. P. (2004) The hydratase activity of malonate semialdehyde decarboxylase: mechanistic and evolutionary implications. *J. Am. Chem. Soc.* **126**, 15658–15659
18. Azurmendi, H. F., Wang, S. C., Massiah, M. A., Poelarends, G. J., Whitman, C. P., and Mildvan, A. S. (2004) The roles of active-site residues in the catalytic mechanism of trans-3-chloroacrylic acid dehalogenase: a kinetic, NMR, and mutational analysis. *Biochemistry*. **43**, 4082–4091
19. Bloom, J., Sun, S., and Al-Abed, Y. (2016) MIF, a controversial cytokine: a review of structural features, challenges, and opportunities for drug development. *Expert Opin. Ther. Targets*. **20**, 1463–1475
20. Flaster, H., Bernhagen, J., Calandra, T., and Bucala, R. (2007) The macrophage migration inhibitory factor-glucocorticoid dyad: regulation of inflammation and immunity. *Mol. Endocrinol.* **21**, 1267–1280
21. Sugimoto, H., Suzuki, M., Nakagawa, A., Tanaka, I., and Nishihira, J. (1996) Crystal structure of macrophage migration inhibitory factor from human lymphocyte at 2.1 Å resolution. *FEBS Lett.* **389**, 145–148
22. Bernhagen, J., Calandra, T., Mitchell, R. A., Martin, S. B., Tracey, K. J., Voelter, W., Manogue, K. R., Cerami, A., and Bucala, R. (1993) MIF is a pituitary-derived cytokine that potentiates lethal endotoxaemia. *Nature*. **365**, 756–759
23. Kleemann, R., Kapurniotu, A., Frank, R. W., Gessner, A., Mischke, R., Flieger, O., Jüttner, S., Brunner, H., and Bernhagen, J. (1998) Disulfide analysis reveals a role for macrophage migration inhibitory factor (MIF) as thiol-protein oxidoreductase. *J. Mol. Biol.* **280**, 85–102
24. Rosengren, E., Bucala, R., Aman, P., Jacobsson, L., Odh, G., Metz, C. N., and Rorsman, H. (1996) The immunoregulatory mediator macrophage migration inhibitory factor (MIF) catalyzes a tautomerization reaction. *Mol. Med.* **2**, 143–149
25. Sparkes, A., De Baetselier, P., Roelants, K., De Trez, C., Magez, S., Van Ginderachter, J. A., Raes, G., Bucala, R., and Stijlemans, B. (2017) The non-mammalian MIF superfamily. *Immunobiology*. **222**, 473–482
26. Fingerle-Rowson, G., Kaleswarapu, D. R., Schlander, C., Kabgani, N., Brocks, T., Reinart, N., Busch, R., Schütz, A., Lue, H., Du, X., Liu, A., Xiong, H., Chen, Y.,

- Nemajerova, A., Hallek, M., Bernhagen, J., Leng, L., and Bucala, R. (2009) A tautomerase-null macrophage migration-inhibitory factor (MIF) gene knock-in mouse model reveals that protein interactions and not enzymatic activity mediate MIF-dependent growth regulation. *Mol. Cell. Biol.* **29**, 1922–1932
27. Poelarends, G. J., Serrano, H., Johnson, W. H., and Whitman, C. P. (2004) Stereospecific alkylation of cis-3-chloroacrylic acid dehalogenase by (R)-oxirane-2-carboxylate: analysis and mechanistic implications. *Biochemistry.* **43**, 7187–7196
28. Almrud, J. J., Poelarends, G. J., Johnson, W. H., Serrano, H., Hackert, M. L., and Whitman, C. P. (2005) Crystal structures of the wild-type, P1A mutant, and inactivated malonate semialdehyde decarboxylase: a structural basis for the decarboxylase and hydratase activities. *Biochemistry.* **44**, 14818–14827
29. Taylor, A. B., Czerwinski, R. M., Johnson, W. H., Whitman, C. P., and Hackert, M. L. (1998) Crystal structure of 4-oxalocrotonate tautomerase inactivated by 2-oxo-3-pentynoate at 2.4 Å resolution: analysis and implications for the mechanism of inactivation and catalysis. *Biochemistry.* **37**, 14692–14700
30. Burks, E. A., Fleming, C. D., Mesecar, A. D., Whitman, C. P., and Pegan, S. D. (2010) Kinetic and structural characterization of a heterohexamer 4-oxalocrotonate tautomerase from *Chloroflexus aurantiacus* J-10-fl: implications for functional and structural diversity in the tautomerase superfamily. *Biochemistry.* **49**, 5016–5027
31. Almrud, J. J., Kern, A. D., Wang, S. C., Czerwinski, R. M., Johnson, W. H., Murzin, A. G., Hackert, M. L., and Whitman, C. P. (2002) The crystal structure of YdcE, a 4-oxalocrotonate tautomerase homologue from *Escherichia coli*, confirms the structural basis for oligomer diversity. *Biochemistry.* **41**, 12010–12024
32. Yu, D., Xu, F., Valiente, J., Wang, S., and Zhan, J. (2013) An indigoidine biosynthetic gene cluster from *Streptomyces chromofuscus* ATCC 49982 contains an unusual IndB homologue. *J. Ind. Microbiol. Biotechnol.* **40**, 159–168
33. Poelarends, G. J., Serrano, H., Person, M. D., Johnson, W. H., and Whitman, C. P. (2008) Characterization of Cg10062 from *Corynebacterium glutamicum*: implications for the evolution of cis-3-chloroacrylic acid dehalogenase activity in the tautomerase superfamily. *Biochemistry.* **47**, 8139–8147
34. Poelarends, G. J., Serrano, H., Person, M. D., Johnson, W. H., Murzin, A. G., and Whitman, C. P. (2004) Cloning, expression, and characterization of a cis-3-chloroacrylic acid dehalogenase: insights into the mechanistic, structural, and evolutionary relationship between isomer-specific 3-chloroacrylic acid dehalogenases. *Biochemistry.* **43**, 759–772
35. Atkinson, H. J., Morris, J. H., Ferrin, T. E., and Babbitt, P. C. (2009) Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PLoS ONE.* **4**, e4345
36. Barber, A. E., and Babbitt, P. C. (2012) Pythoscape: a framework for generation of large protein similarity networks. *Bioinformatics.* **28**, 2845–2846
37. Eddy, S. R. (1998) Profile hidden Markov models. *Bioinformatics.* **14**, 755–763
38. Atkinson, H. J., and Babbitt, P. C. (2009) An atlas of the thioredoxin fold class reveals the complexity of function-enabling adaptations. *PLoS Comput Biol.* **5**, e1000541
39. Brown, S. D., and Babbitt, P. C. (2012) Inference of functional properties from large-scale analysis of enzyme superfamilies. *J. Biol. Chem.* **287**, 35–42

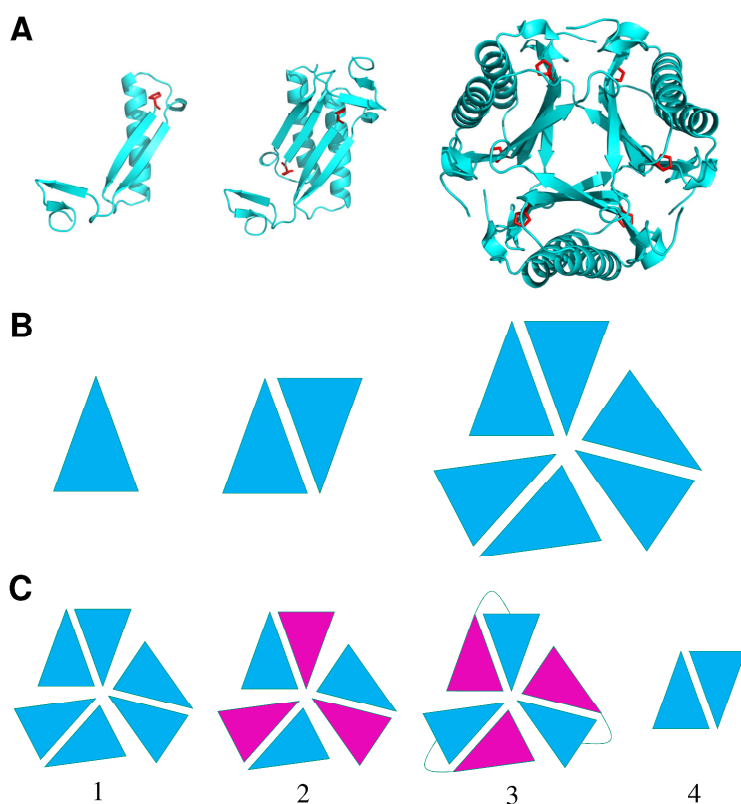
40. Uberto, R., and Moomaw, E. W. (2013) Protein similarity networks reveal relationships among sequence, structure, and function within the Cupin superfamily. *PLoS ONE*. **8**, e74477
41. Mashiyama, S. T., Malabanan, M. M., Akiva, E., Bhosle, R., Branch, M. C., Hillerich, B., Jagessar, K., Kim, J., Patskovsky, Y., Seidel, R. D., Stead, M., Toro, R., Vetting, M. W., Almo, S. C., Armstrong, R. N., and Babbitt, P. C. (2014) Large-scale determination of sequence, structure, and function relationships in cytosolic glutathione transferases across the biosphere. *PLoS Biol.* **12**, e1001843
42. van der Meer, J.-Y., Poddar, H., Baas, B.-J., Miao, Y., Rahimi, M., Kunzendorf, A., van Merkerk, R., Tepper, P. G., Geertsema, E. M., Thunnissen, A.-M. W. H., Quax, W. J., and Poelarends, G. J. (2016) Using mutability landscapes of a promiscuous tautomerase to guide the engineering of enantioselective Michaelases. *Nat Commun.* **7**, 10911
43. Czerwinski, R. M., Harris, T. K., Johnson, W. H., Legler, P. M., Stivers, J. T., Mildvan, A. S., and Whitman, C. P. (1999) Effects of mutations of the active site arginine residues in 4-oxalocrotonate tautomerase on the pKa values of active site residues and on the pH dependence of catalysis. *Biochemistry*. **38**, 12358–12366
44. Czerwinski, R. M., Harris, T. K., Massiah, M. A., Mildvan, A. S., and Whitman, C. P. (2001) The structural basis for the perturbed pKa of the catalytic base in 4-oxalocrotonate tautomerase: kinetic and structural effects of mutations of Phe-50. *Biochemistry*. **40**, 1984–1995
45. Harris, T. K., Czerwinski, R. M., Johnson, W. H., Legler, P. M., Abeygunawardana, C., Massiah, M. A., Stivers, J. T., Whitman, C. P., and Mildvan, A. S. (1999) Kinetic, stereochemical, and structural effects of mutations of the active site arginine residues in 4-oxalocrotonate tautomerase. *Biochemistry*. **38**, 12343–12357
46. Poelarends, G. J., Almrud, J. J., Serrano, H., Darty, J. E., Johnson, W. H., Hackert, M. L., and Whitman, C. P. (2006) Evolution of enzymatic activity in the tautomerase superfamily: mechanistic and structural consequences of the L8R mutation in 4-oxalocrotonate tautomerase. *Biochemistry*. **45**, 7700–7708
47. Huddleston, J. P., Burks, E. A., and Whitman, C. P. (2014) Identification and characterization of new family members in the tautomerase superfamily: analysis and implications. *Arch. Biochem. Biophys.* **564**, 189–196
48. Poelarends, G. J., Johnson, W. H., Murzin, A. G., and Whitman, C. P. (2003) Mechanistic characterization of a bacterial malonate semialdehyde decarboxylase: identification of a new activity on the tautomerase superfamily. *J. Biol. Chem.* **278**, 48674–48683
49. Hirel, P. H., Schmitter, M. J., Dessen, P., Fayat, G., and Blanquet, S. (1989) Extent of N-terminal methionine excision from *Escherichia coli* proteins is governed by the side-chain length of the penultimate amino acid. *Proc. Natl. Acad. Sci. U.S.A.* **86**, 8247–8251
50. Almrud, J. J., Dasgupta, R., Czerwinski, R. M., Kern, A. D., Hackert, M. L., and Whitman, C. P. (2010) Kinetic and structural characterization of DmpI from *Helicobacter pylori* and *Archaeoglobus fulgidus*, two 4-oxalocrotonate tautomerase family members. *Bioorg. Chem.* **38**, 252–259
51. Nishihira, J., Fujinaga, M., Kuriyama, T., Suzuki, M., Sugimoto, H., Nakagawa, A., Tanaka, I., and Sakai, M. (1998) Molecular cloning of human D-dopachrome

- tautomerase cDNA: N-terminal proline is essential for enzyme activation. *Biochem. Biophys. Res. Commun.* **243**, 538–544
52. Merk, M., Zierow, S., Leng, L., Das, R., Du, X., Schulte, W., Fan, J., Lue, H., Chen, Y., Xiong, H., Chagnon, F., Bernhagen, J., Lolis, E., Mor, G., Lesur, O., and Bucala, R. (2011) The D-dopachrome tautomerase (DDT) gene product is a cytokine and functional homolog of macrophage migration inhibitory factor (MIF). *Proc. Natl. Acad. Sci. U.S.A.* **108**, E577–85
  53. Merk, M., Mitchell, R. A., Endres, S., and Bucala, R. (2012) D-dopachrome tautomerase (D-DT or MIF-2): doubling the MIF cytokine family. *Cytokine*. **59**, 10–17
  54. Panstruga, R., Baumgarten, K., and Bernhagen, J. (2015) Phylogeny and evolution of plant macrophage migration inhibitory factor/D-dopachrome tautomerase-like proteins. *BMC Evol. Biol.* **15**, 64
  55. Chauhan, N., Sharma, R., and Hoti, S. L. (2015) Identification and biochemical characterization of macrophage migration inhibitory factor-2 (MIF-2) homologue of human lymphatic filarial parasite, *Wuchereria bancrofti*. *Acta Trop.* **142**, 71–78
  56. Whitman, C. P., Hajipour, G., Watson, R. J., J. J., Bembenek, M. E., and Stolowich, N. J. (1992) Stereospecific ketonization of 2-hydroxymuconate by 4-oxalocrotonate tautomerase and 5-carboxymethyl-2-hydroxymuconate isomerase. *J. Am. Chem. Soc.* **114**, 10104–10110
  57. Poelarends, G. J., Johnson, W. H., Serrano, H., and Whitman, C. P. (2007) Phenylpyruvate tautomerase activity of trans-3-chloroacrylic acid dehalogenase: evidence for an enol intermediate in the dehalogenase reaction? *Biochemistry*. **46**, 9596–9604
  58. LeVieux, J. A., Baas, B.-J., Kaoud, T. S., Davidson, R., Babbitt, P. C., Zhang, Y. J., and Whitman, C. P. (2017) Kinetic and structural characterization of a cis-3-Chloroacrylic acid dehalogenase homologue in *Pseudomonas* sp. UW4: A potential step between subgroups in the tautomerase superfamily. *Arch. Biochem. Biophys.* **636**, 50–56
  59. Chothia, C., and Lesk, A. M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823–826
  60. Xu, J., and Zhang, Y. (2010) How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics*. **26**, 889–895
  61. Schnoes, A. M., Brown, S. D., Dodevski, I., and Babbitt, P. C. (2009) Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput Biol.* **5**, e1000605
  62. Finn, R. D., Attwood, T. K., Babbitt, P. C., Bateman, A., Bork, P., Bridge, A. J., Chang, H.-Y., Dosztányi, Z., El-Gebali, S., Fraser, M., Gough, J., Haft, D., Holliday, G. L., Huang, H., Huang, X., Letunic, I., Lopez, R., Lu, S., Marchler-Bauer, A., Mi, H., Mistry, J., Natale, D. A., Necci, M., Nuka, G., Orengo, C. A., Park, Y., Pesseat, S., Piovesan, D., Potter, S. C., Rawlings, N. D., Redaschi, N., Richardson, L., Rivoire, C., Sangrador-Vegas, A., Sigrist, C., Sillitoe, I., Smithers, B., Squizzato, S., Sutton, G., Thanki, N., Thomas, P. D., Tosatto, S. C. E., Wu, C. H., Xenarios, I., Yeh, L.-S., Young, S.-Y., and Mitchell, A. L. (2017) InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res.* **45**, D190–D199
  63. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and

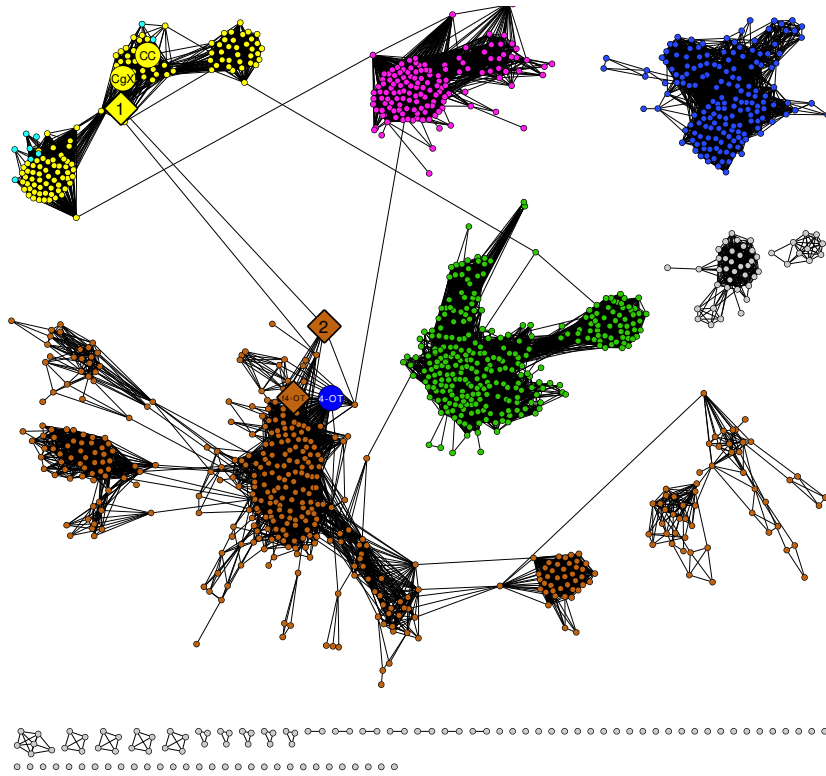


- Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402
64. Li, W., and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics.* **22**, 1658–1659
65. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504
66. Zhang, Y. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309
67. Finn, R. D., Coghill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate, J., and Bateman, A. (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44**, D279–85
68. Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., and Higgins, D. G. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539–539
69. Eddy, S. R. (2011) Accelerated Profile HMM Searches. *PLoS Comput Biol.* **7**, e1002195
70. Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004) UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem.* **25**, 1605–1612
71. Edgar, R. C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797
72. Ronquist, F., and Huelsenbeck, J. P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics.* **19**, 1572–1574

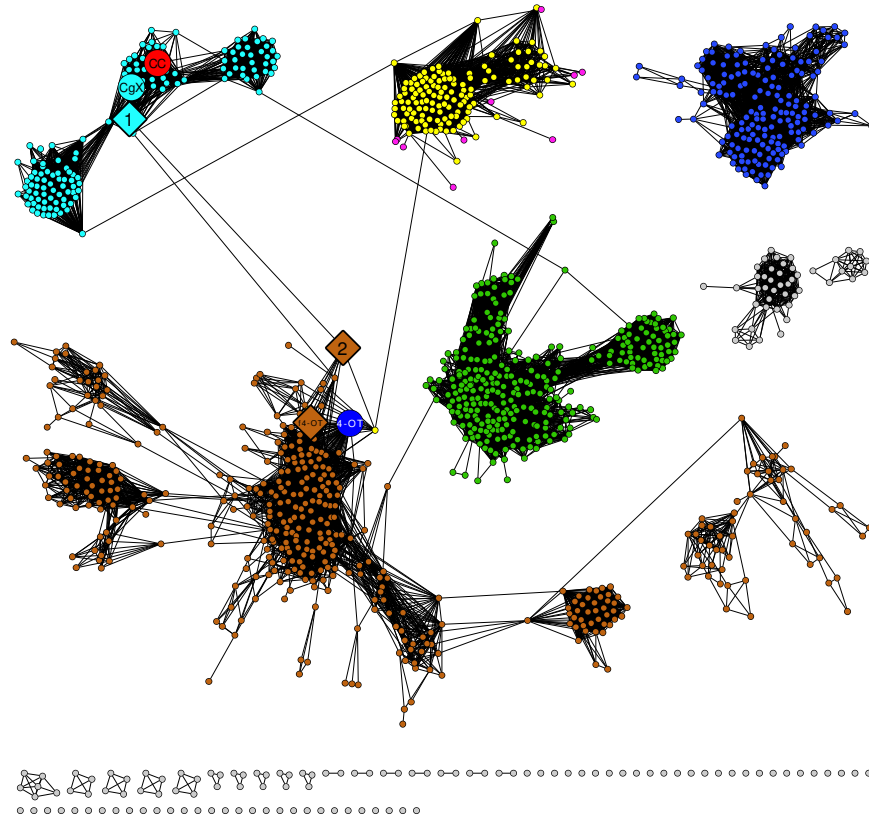
## Appendix A. Supplemental Information for Chapter 2



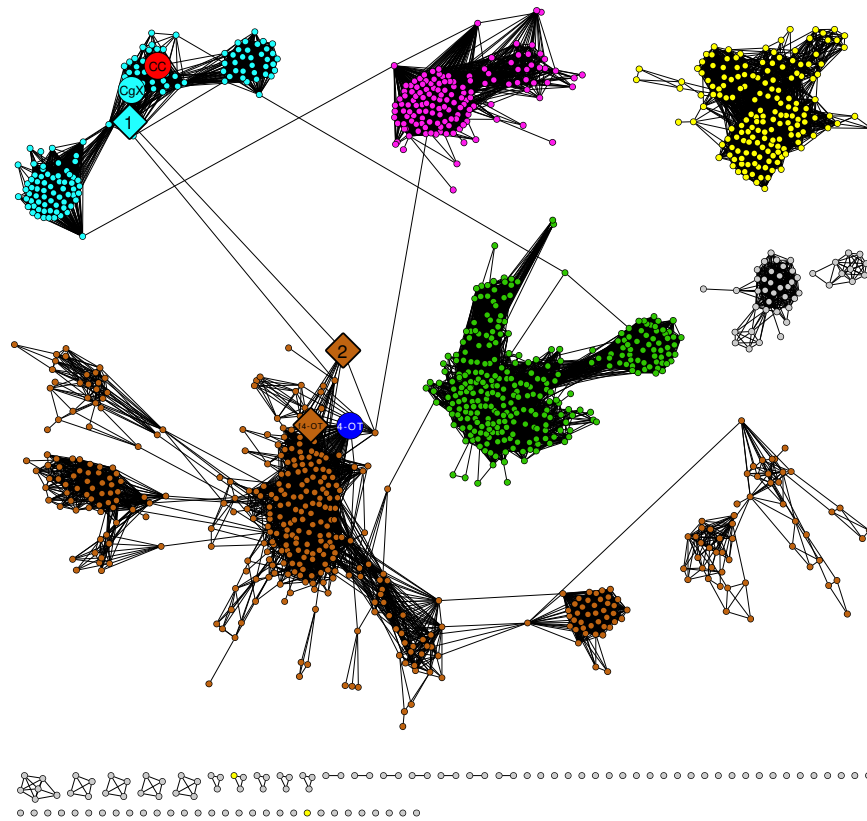
**Figure A.1 Oligomeric organization in the TSF.** A. The three levels of structure for the founder 4-OT. The 62-amino acid monomer (left) showing the  $\beta$ - $\alpha$ - $\beta$ -motif, the signature fold of the tautomerase superfamily. Monomers stack in an antiparallel fashion to yield a dimer (middle), which then make up the active homohexamer (right). B. Schematic representation of the same three structures. C. The four distinct patterns observed for the experimentally characterized members of the TSF. 1. Homohexamer, which includes the ‘short’ tautomerasases such as the founder 4-OT. 2. Heterohexamer, which includes the hh4-OT and CaaD. Each heterohexamer is composed of three  $\alpha$ , $\beta$  dimers, where the  $\alpha$ - and  $\beta$ -subunit fold into a  $\beta$ - $\alpha$ - $\beta$ -motif, but are not identical in sequence. 3. Trimer, which includes ‘long’ members of the *cis*-CaaD-, MIF-, MSAD- and CHMI-subgroups. Note that the monomers in each of these trimers are identical in sequence, but the two  $\beta$ - $\alpha$ - $\beta$ -motifs within each monomer are not. 4. Homodimer, which includes YdcE, a 4-OT homolog from *E. coli*.



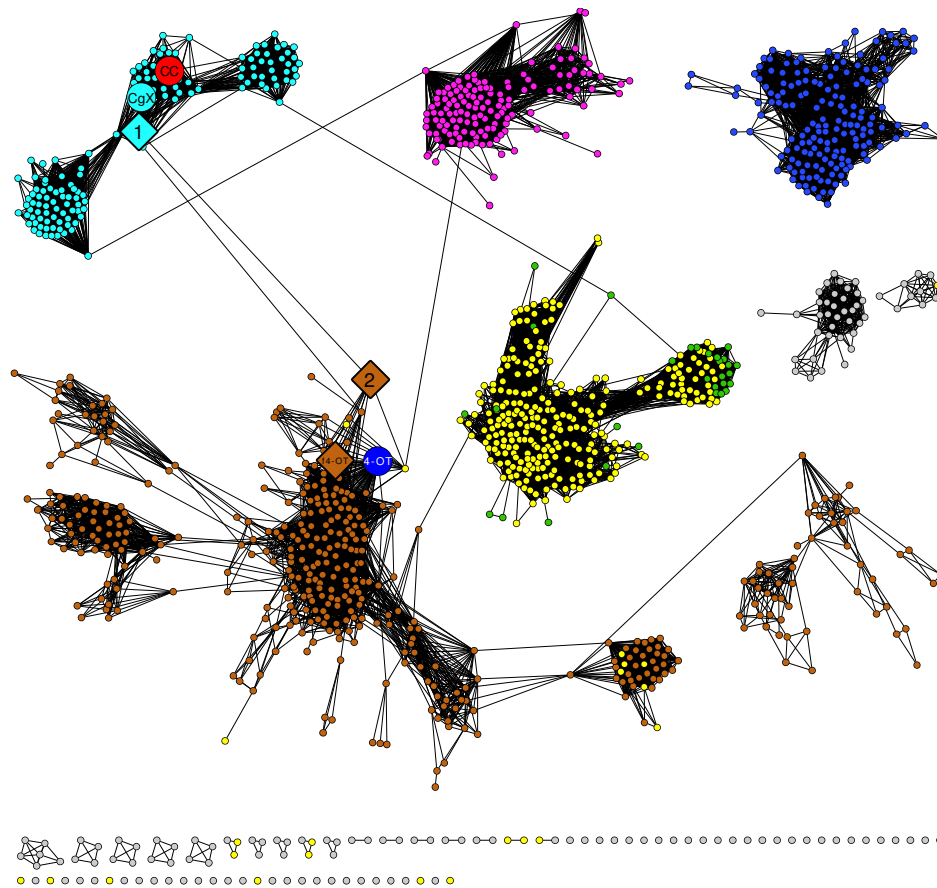
**Figure A.2 Mapping of the Level 1 *cis*-CaaD HMM to the Level 1 *cis*-CaaD subgroup.** Details and color for this SSN are as in Figure 2.2 except that yellow designates each representative node that contains at least one sequence for which its cognate HMM hits this subgroup at an HMM hit score of  $1e^{-20}$ . Nodes that retain the original cyan color of the *cis*-CaaD subgroup were missed by the HMM trained on this subgroup. The large labeled nodes indicate representative nodes that link the 4-OT and *cis*-CaaD subgroups as described in the section “*Linkers*” between *cis*-CaaD and 4-OT subgroup identify structural transitions between them.” The representative node of founder 4-OT was colored dark blue for consistency with Figure 2.5, and was not matched by any subgroup HMM other than that generated for the 4-OT level 2 subgroup 1 (see Figure A.7).



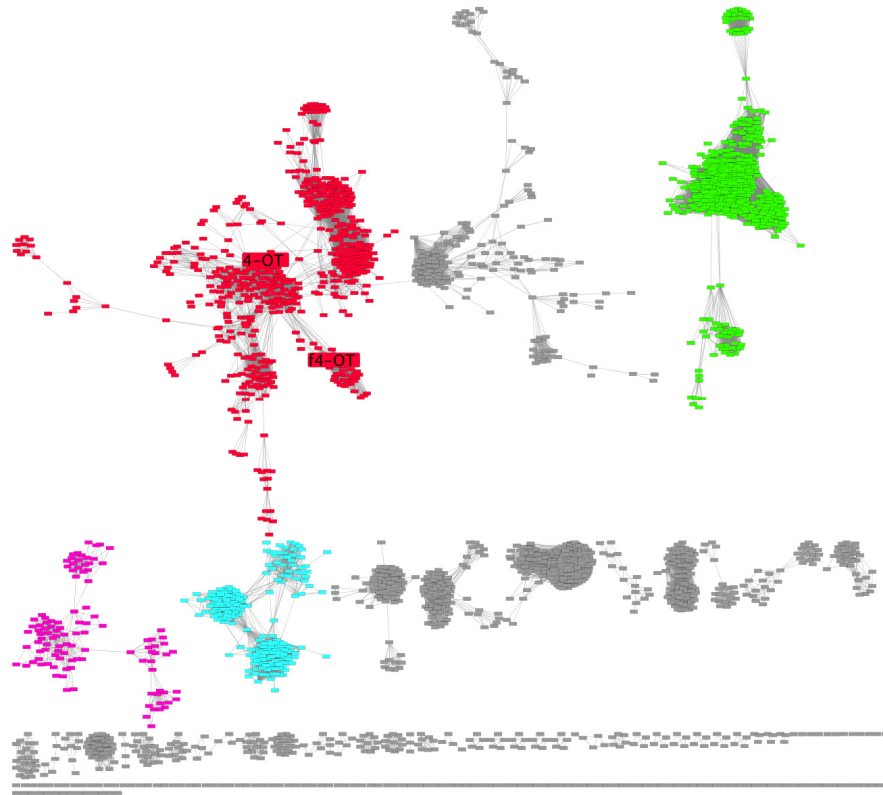
**Figure A.3 Mapping of the Level 1 MSAD HMM to the Level 1 MSAD subgroup.** Details and color for this SSN are as in Figure 2.2 except that yellow designates each representative node that contains at least one sequence for which its cognate HMM hits this subgroup or other subgroups at an HMM hit score of  $1e^{-14}$ . Nodes that retain the original magenta color of the MSAD subgroup were missed by the HMM trained on this subgroup. The large labeled nodes indicate representative nodes that link the 4-OT and *cis*-CaaD subgroups as described in the section ““Linkers” between *cis*-CaaD and 4-OT subgroup identify structural transitions between them.” The representative node of founder 4-OT was colored dark blue for consistency with Figure 2.5, and was not matched by any subgroup HMM other than that generated for the 4-OT level 2 subgroup 1 (see Figure A.7).



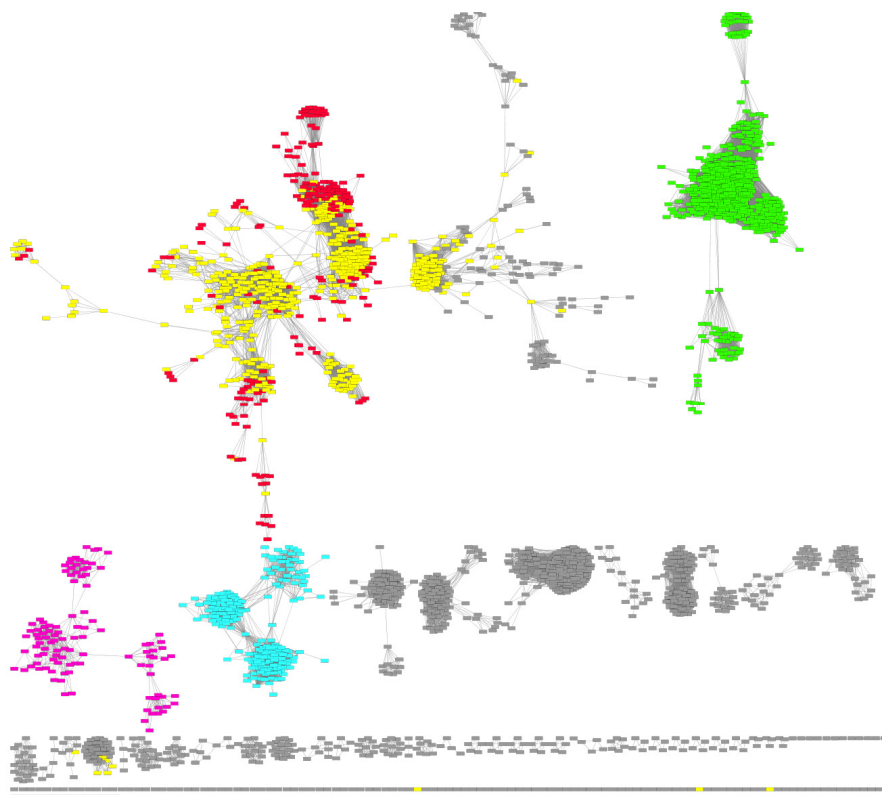
**Figure A.4 Mapping of the Level 1 CHMI HMM to the Level 1 CHMI subgroup.** Details and color for this SSN are as in Figure 2.2 except that yellow designates each representative node that contains at least one sequence for which its cognate HMM hits this subgroup at an HMM hit score of  $1e^{-10}$ . Nodes that retain the original dark blue color of the CHMI subgroup (only the representative founder 4-OT node) were missed by the HMM trained on this subgroup. The large labeled nodes indicate representative nodes that link the 4-OT and *cis*-CaaD subgroups as described in the section “*Linkers*” between *cis*-CaaD and 4-OT subgroup identify structural transitions between them.” The representative node of founder 4-OT was colored dark blue for consistency with Figure 2.5, and was not matched by any subgroup HMM other than that generated for the 4-OT level 2 subgroup 1 (see Figure A.7).



**Figure A.5 Mapping of the Level 1 MIF HMM to the Level 1 MIF subgroup.** Details and color for this SSN are as in Figure 2.2 except that yellow designates each representative node that contains at least one sequence for which its cognate HMM hits this subgroup or other subgroups at an HMM hit score of  $1e^{-11}$ . Nodes that retain the original green color of the MIF subgroup were missed by the HMM trained on this subgroup. The large labeled nodes indicate representative nodes that link the 4-OT and *cis*-CaaD subgroups as described in the section “*Linkers between cis-CaaD and 4-OT subgroup identify structural transitions between them.*” The representative node of founder 4-OT was colored dark blue for consistency with Figure 2.5, and was not matched by any subgroup HMM other than that generated for the 4-OT level 2 subgroup 1 (see Figure A.7).

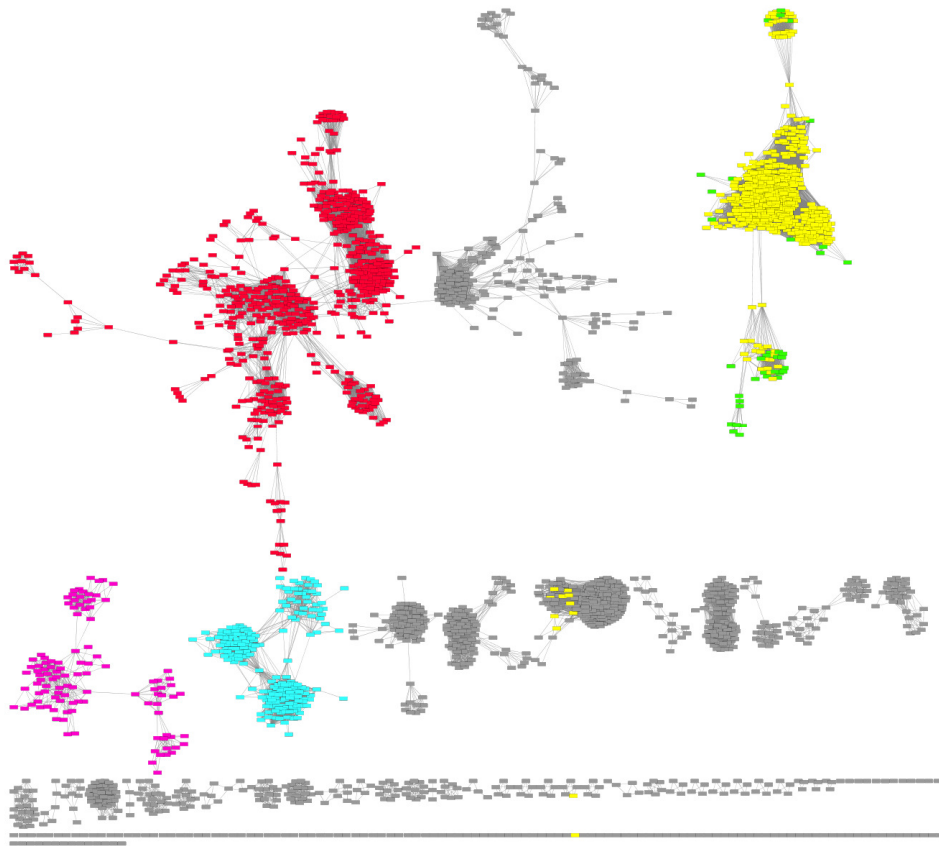


**Figure A.6 90% sequence identity per node network of Level 2 subgroups of the Level 1 4-OT subgroup.** 2,580 network nodes represent 4,530 4-OT like proteins colored by level-2 subgroups. The threshold for drawing edges between representative nodes is  $1e^{-18}$ . Gray nodes designate nodes that have not been assigned to a level-2 subgroup. HMMs have been generated for subgroup 1 (red), subgroup 2 (green), subgroup 3 (magenta) and subgroup 4 (cyan). The founder 4-OT and the linker Fused-4-OT proteins belong to the Level 2 subgroup 1, as indicated by the labeled large nodes.

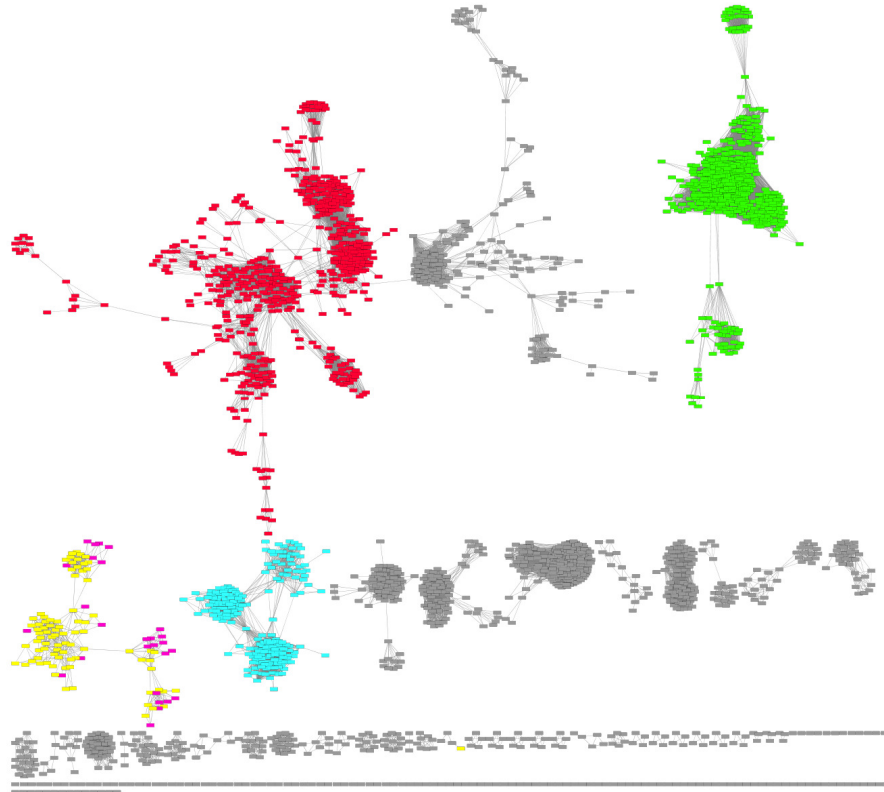


**Figure A.7 HMM mapping of the Level 2 subgroup 1 to the Level 1 4-OT subgroup.** Details and color as in Figure A.6 except that yellow designates each representative node that contains at least one sequence for which its cognate HMM hits this subgroup at an HMM hit score of  $1e^{-24}$ . Nodes that retain the original red color of Level 2 subgroup 1 were missed by the HMM trained on this subgroup. Yellow nodes that are not part of this subgroup are also matched by this HMM at the HMM hit score.

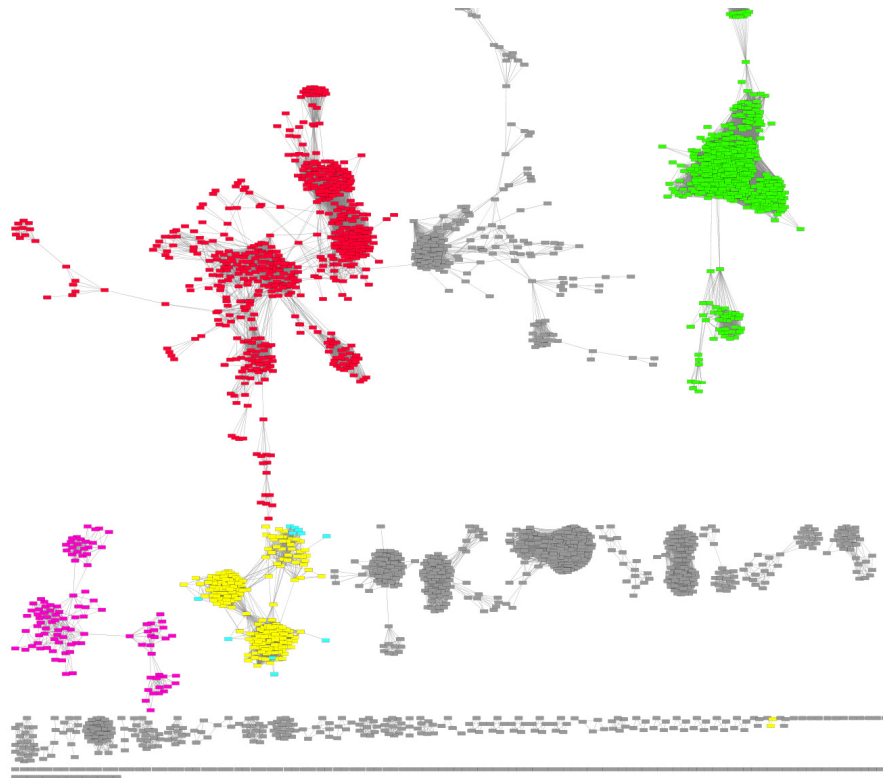




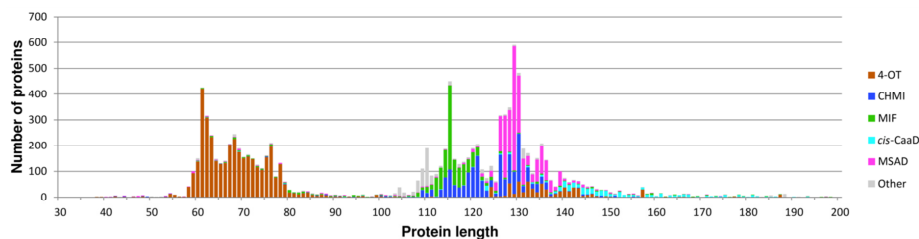
**Figure A.8 HMM mapping of the Level 2 subgroup 2 to the Level 1 4-OT subgroup.** Details and color as in Figure A.6 except that yellow designates each representative node that contains at least one sequence for which its cognate HMM hits this subgroup with an HMM hit score of  $1e^{-23}$ . Nodes that retain the original green color of Level 2 subgroup 2 were missed by the HMM trained on this subgroup. Yellow nodes that are not part of this subgroup are also matched by this HMM at the HMM hit score.



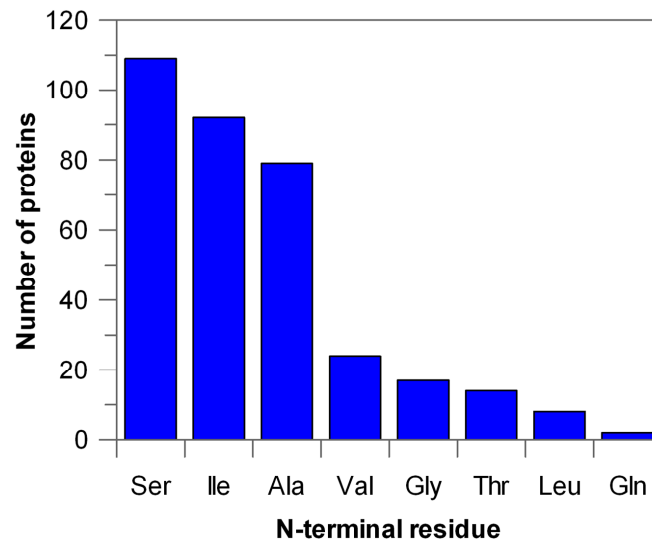
**Figure A.9 HMM mapping of the Level 2 subgroup 3 to the Level 1 4-OT subgroup.** Details and color as in Figure A.6 except that yellow designates each representative node that contains at least one sequence for which its cognate HMM hits this with an HMM hit score of  $1e^{-25}$ . Nodes that retain the original magenta color of Level 2 subgroup 3 were missed by the HMM trained on this subgroup. Yellow nodes that are not part of this subgroup are matched by this HMM at the HMM hit score.



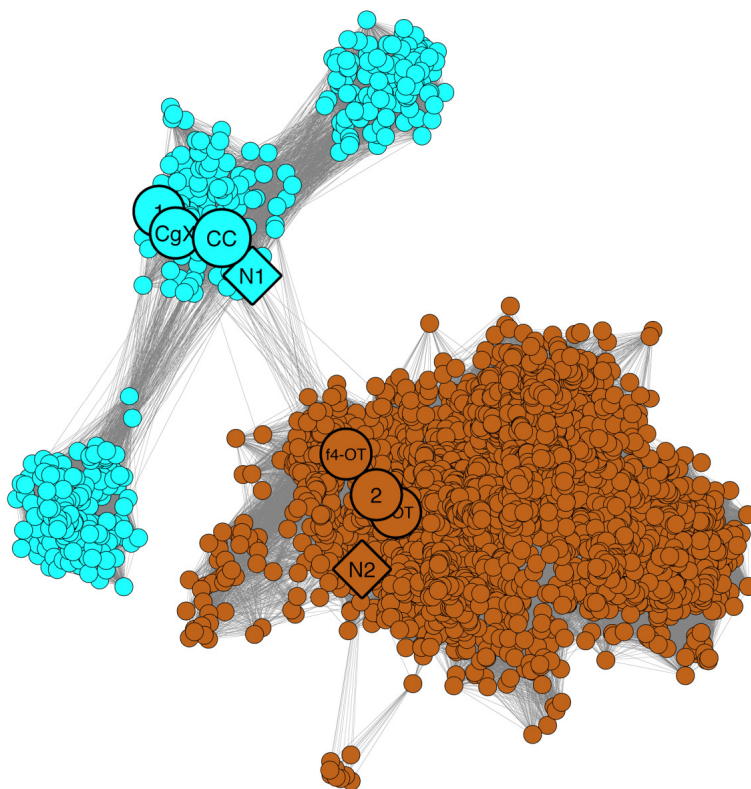
**Figure A.10 HMM mapping of the Level 2 subgroup 4 to the Level 1 4-OT subgroup.** Details and color as in Figure A.6 except that yellow designates each representative node that contains at least one sequence for which its cognate HMM hits this subgroup with an HMM hit score of  $1e^{-22}$ . Nodes that retain the original cyan color of Level 2 subgroup 4 were missed by the HMM trained on this subgroup. Yellow nodes that are not part of this subgroup are also matched by this HMM at the HMM hit score.



**Figure A.11 Length histogram of 11,395 non-redundant protein sequences in the TSF.** The histogram is composed of stacked bars colored by subgroups as in Figure 2.2. A complex pattern in sequence lengths per subgroup is apparent. Sequences between 58-84 residues in length, which fold into a single  $\beta$ - $\alpha$ - $\beta$ -unit, largely belong to the 4-OT-subgroup. Sequences between 110-150 residues in length are found in the other subgroups and fold into two fused  $\beta$ - $\alpha$ - $\beta$ -units. There is “crosstalk” between these two populations, most notably the Fused 4-OTs. These proteins are part of the 4-OT subgroup in the SSN, but have a length similar to that of two fused  $\beta$ - $\alpha$ - $\beta$ -units. The histogram shows sequences up to 200 residues in length, but there are longer members. These may represent proteins with a TSF-like domain, e.g. the 1357-residue indigoidine synthase IndC from *Streptomyces clavuligerus* that has a 76-residue C-terminal 4-OT-like domain.



**Figure A.12 Non-Pro-1 Frequencies in the TSF.** Histogram showing the frequency of residues in place of Pro-1 in the curated set of bacterial non-Pro-1 sequences (346 in total). Serine (109), isoleucine (92), alanine (80), valine (24), glycine (17), threonine (14), leucine (8), and glutamine (2).



**Figure A.13 Linker control network.** One sequence per node network of 2,761 non-redundant sequences expanded from the 90% representative network of the 4-OT and *cis*-CaaD subgroups. The threshold for drawing edges is  $1e^{-11}$ . The positions of the linker nodes are highlighted as shown in Figure 2.5, along with the positions of two additional linker sequences, N1 and N2. These latter nodes refer to the second pair of representative nodes besides Linker 1 and Linker 2 that link the 4-OT and *cis*-CaaD subgroups as shown in Figure 2.5. In this SSN, 48 edges link 31 nodes from the founder 4-OT subgroup to 17 nodes in the *cis*-CaaD subgroup.

Figure A.14 MSA of sequences used to calculate the phylogenetic tree.

```

                10      20      30      40      50      60      70
cis-CaaD/1-149  PVYMYVVSQDR LTPSAKHAVAKA I TDAHRGLTGTQHFLAQVNFNEQPA GNVFLGGV ----- QQGG-DT I
CgX/1-148      P TYTCWSQR I R I SREAKQR I AEA I TDAHHELAHAPKYLVQV I FNEVEPDSYF I AAQ ----- SASE-NH I
Linker_1/2-113 PLYECQTVKGT L SERQRQSLAES I TS IHTRE TGAPASYVHV L FKELEPGSA FTAGQ ----- PAAT --- A
Linker_2/2-124 P YV T I SA TEG -LSAEKKKQL LERSSDAVVQS I GAP LASVRVMLHELPGGHYLNAGQ ----- FNT PGL
Fused_4-OT/2-128 P TLEVFLPAG -HDDARKAE L I ARLTGA TVDS I GAP IESVRV L L TELP A TH I GLGGR --- SAADGAPPSLP
A0A125QJV7/1-135 P LY I CNAKAGAVPENAKAK I AEDVTR I HCEV TDAPPT FVHV F ----- HFADGP ----- MPP LGDKQV
A0A0Q6VD17/1-128 P LV I LTQAGMLDQNAKVE LAEKL T ALHCEYAGVPANWH I I FQDYR TGDGF I AGK ----- PAAT --- T
A0A0C4YAG5/1-122 P V I VCECKVG - I EAGVKAK I ASEFTSA I RE I I LSP LDL I SVV FHES TPENTYRSGE ----- P TSE --- T
X0QBL2/1-122   P V I QCHNRAG - LDLETKAQLAE I TRDVHE I VKSPNDL I SV I FSDLPAES TYRNGQ ----- ATNE --- T
A0A060ZNL3/1-122 P V I QCENRAG - FAPDVKAKLADE I TAVVRDV I KSPMDL I SV I FHDLPPESTYRSGA ----- P TDE --- T
A0A149PHW3/1-122 P I I VCNTRAG - LDVDVKKR I AKA I T I AVNET I KSPLE I I SVV FNDL SSES SYV GGE ----- PGGD --- T
A0A0Q2/1-122   P I I LNVQ I LQG - HSAAQKAA L KAASNAVVES I AAP LPSVR I VLQEVPAEHV I VAGE ----- I GK RMA
A0A149PTP3/1-123 P TLEVYLPQG - YADERKAQL I ESLTEA TVQA I GAPAESVRV L LSE LGA TNAGLGGK ----- L TPAALP
A0A0S1Y110/1-123 P T I HAH I TAG -PSAQKQSAL LQAASQAVVES LGAP LRSVRLMLHVDVARED I VVGGE ----- VGRESV
A0A157SPE8/1-123 P I L N I Q I MQG -HTDSEK T ALLENA TRAVEAS I GAPVQS I R I V LEEVDARNV I VAGK ----- LGHPMA
H0FC32/1-123   P I I LNVQ I LQG - HSAAQKAA L KAASNAVVES I AAP LPSVR I VLQEVPAEHV I VAGE ----- I GK RMA
A0A158M531/1-123 P I LNVQ I MQG -YAPSAK T ALLKLSQAVVDS I GAPVASVR I TLAEVDAGHV I VAGE ----- LGKDMA
A0A063UWJ0/1-123 P L LNVQ I MQG -HQPAQKSA L LAAACQAVVDS I GAP LASVRV L LEEVPAAHV I VAGE ----- LGRQMA
A0A142YAD3/1-125 P M Y T I V I QSGS VDE TAKAS LAAE I T ALHVELSAV PKDWHV V FQEYAPGSGFNAGE ----- AG-PLV
A0A0M4QA63/1-129 P Y Y R F T V P T GNA T LAHRAEVA AAV TRVHSEV TGAPARYVHCTF I EVPPGSV FVAGE ----- PVGE --- P
L8F8C4/1-130   P V Y T C T T A V E T L T P S S K A A L A T E I T R I H S E V N H V P S T Y I N V V F Q E L P A E D V Y T D A R ----- P A Q P --- L
B1FFZ6/1-127   P TLEVFLPAG -HDDARKAE L I ARLSGA TVDA I GAP IESVRV L L TELP A TH I GLGGR --- TAADGAPPSLP
A0A0F5K0U6/1-130 P T L Q I H L A Q G - H D A S Q K A A L I A A L S H A T V E A L T V P I E S V R I M I E D Y A A I D H G A A G N A V P A A A S R A E A P S P
LOIZU6/1-131   P I Y T C T T T E S T L T T D T R T A L A G E I T R L H S A I N H V P S T Y N N V V F H E L P A D A I Y T D G R ----- P A A P --- V
A0A101RJK5/1-132 P V Y G V T T T R G I V S D E Q K A D L A A E I T R I H S S V T G A P S S F V H V V F T E L P E A N V F T D S A ----- P S R P --- L
A0A101QRI2/1-132 P I Y T C T T A Q G T L K A G A K G E L A A E I T R I H A E I N H V P P A Y V N N V F S E L P Q D S V Y V G G E ----- P G A P --- L
A0A0U3HBV1/1-132 P I Y T C T C V Q G S L P S D V K D G L A A E I T R I H A D I N G V P P D Y N N V V F T E T P R E N V Y V G G R ----- P G T P --- L
W9AW45/1-134   P V Y T C T T V R S L S A D T K A D L A A E V T R I H S M I N H V P G T Y N N V V F H E L A P D E V Y T G G H ----- P A Q P --- L
A0A0Q7FJ36/1-134 P T I E A F I A E G - Y S H R Q K H D L I E A M T H A V D A I D A P I D S I R V I L N E V A T H D A G I G G K ----- T G A A I S
D6TDA1/1-136   A V Y T C I T Q E G A L S A E Q R A D L A Q E I T R S H V A I S G E P A S F V R V I F E T V S P N S T F S G G K ----- P A V N --- A
R4X1K7/1-148   P T Y H V S A T E G L L D D K A R A R L A N E I T R A H S L A T G A Q S F F A Q V L F H E T P K R F H F M G G K ----- P V D S - E Q V
A0QSU8/1-138   P V Y T V T M S R G T L N G E T K A A L A A E I T T I H S A V N H V P S T Y N N V L F N E L A P S N V Y T D G K ----- P A H P --- L
W4JVR2/1-138   P L Y I L Y Y H A G S L S F E Q Q K T I A K E V T D A H C A V T G A G P Y F V K V A F Q A C E R G D F Y T A G D ----- I D Q - K H V
A0A094IHN9/1-139 P T F V V T A T A G C L S A D K K A A I A N A L V T I H S T E N G V P R W C V Q V I F H E I S A G N H F I G D K ----- F A P A - D Q L
A0A0R3F487/1-141 P L Y Q C I T R A E S L T D E V R Q T I A Q E F T R I H C E I T G A P A V F V H V V F N E Y Q P G H H Y L A G K ----- P E S - D T T
I2AHZ4/1-141   P V Y Q C Y S P P G L L S E S A K A R I A D E I T T I H T S A T G A P E L F V N V L F H E I P V G D C F V G R K ----- Q A S H S
A0A0P8AZP5/1-142 P T Y T V T V A N L S L A E Q K S Q I A E A I T A A H N A Q T G A P R F F A Q V L F Y A A N E G D H F V G G R ----- V N T A - P Q V
A0A0D6IB59/1-142 P L Y Q I D T V K G R L T P S V K A E I A N K V T D I H C Q L T G A P D T F V N V V F R E Y T E G D C F V A R K ----- P E G R --- S
W9GWJ5/1-143   P T Y A F S T A K E - L T A E Q R A K L V E S V T S I H Q V E A A P R Y F V Q V I F H K I E P G S M F I G G E ----- A A S P - D H V
A0A0M2XHR9/1-143 P T Y T V R T S T G R L G P D A R A S L A H A I T S A H T V A T G A P G F F A Q V V F E E V D T D R H F V G G T ----- P I A D - E L V
A0A0S2PHN2/1-143 P I Y Q C S A P V G L L T D D M K A A V A T A I T D A H V E A T G A P K A F V H V F F H E L P P G I A Y S A G E ----- L D T D I S
A0A0E4GVS2/1-143 P V Y Q C V S S A G L V S A D A R A R I A Q D I T R L H C E V T G A P A A F V N V L F S E Y V S G E L F T G G Q ----- P S Q N --- S
M1NLA4/1-144   P S Y A V S S R A G L I D Q E R R A A V A D L L T T L H R D I A V A P R Y L V Q V I F N D L D A G A L F L A G R ----- E A P E - G H V
R7XJL4/1-144   P T Y V C W T K T G R L S S E Q R A Q I A K S V T E I H H E V G R A P R Y F V Q V I F N E L G A Q S H F I G G T ----- E A S V - D Q I
A0A0Q5VHB9/1-144 P T Y T C W S E S G L V G P E Q K Q A V A T A L T E I H H E V A V A P R Y F V Q V I F T E L V P G S V F L A G Q ----- P A T R - G H V
A0A0T1WIY8/1-145 P T Y T C W S K T G V V A T E A R A R I A A A L T R I H Y D V A A G P R Y F V Q V I F T E L D P D S L F I G G K ----- P V T A - E H V
A0A0P9CHA8/1-145 P T Y I L K S S K L K I N K T K R D L L A K G I T R I H S K V T G A N S Y F A Q V I F Q E N I N G S H Y M G G K ----- P V Q S - K E V
A0A0B6S9Z8/1-145 P T Y V V S A A A G R F S Q E E K E G I A R G I T Q A H G R A T G A Q G F F A Q V I F H W I A A D D H F V G G R ----- V H D A - D H V
W6X611/1-145   P T Y T V M A A A G R L T D P Q K R E I A R D I T R V H S E A T G A Q G F F A Q V I F Q A I P A G D H F L G G A ----- P L V S - D Q L
F2LFU5/1-145   P T Y V V S A A A G R L D A Q Q K A R I A A G I T E A H S A Q T G A Q G F F A Q V I F V A I A E G D H F L G G R ----- P L K S - D Q I
A0A0Q7D9D7/1-146 P T Y V V T A P Q G R L S A Q Q K E R I G A D I T R V H C T V A S A P A Y F A Q V I F N D V P A G N Y F V G G K ----- L L Q R S D H V
A0A077K980/1-147 P T Y I V S A T Q N L L S P E E K K K V A K E V T R A H S Q A T G A Q G F F A Q V I F S E I P S G S H F M G G V ----- E I S A - K Q I
A0A126YPH4/1-147 P T Y I V S T A A D R F N D E L K H R I A D G I T K A H S R A T G A Q G F F A Q V I F N E I P A G N H F I G G S ----- P L K A - E Q A
I3U9B8/1-127   P I I L Q V Q V T A G - R S Q Q K T A F L Q N A T K V I E Q T L N A A L P S I R I S L Q E V E Q Q D S I V A G Q ----- I G A E F V
K8P2T6/1-148   P T Y T V T T A N L S L S A A Q K G Q I A E A I T A A H H A A T G A P A F F A Q V I F R A L E N G E H F I G G K ----- P N A H - P H V
E6W3C6/1-148   P T Y T V T A P S G R L N A E Q K Q N L A T A I T R A H H D I T G A P T Y F A Q V I F V E V Q P G N Y F V G G A ----- P L A H - D Q I
A0A0D1NR67/1-148 P T Y T C T S A E G R L S A E Q K S R I A S E I T R I H A E V T G A P S Y F A Q V I F S E V A S G N W F M G G V ----- P V A H - D H I
A0A0S7Z4Z0/1-149 P T Y H C S A R M G L D A E R K A R I A R A V T L A H A E T G A P P H L A Q V L F R D V P P Q D H Y V G G A ----- L L E H - D H V
A0A0G9H7B6/1-149 P T Y I V R S T L P S L T A P T K Q R I A Q A I T A A H A D I T G A N T F F A Q V V F D H A P G D D W F I G G V ----- P V E D - A T L
A0A089UPG7/1-151 P T Y Y C H L P K N R V S D V D K H R L A H A I T A R H T E A T G A P S W F V Q V I E E D D D K I R Y L G G E ----- P A G E - H I
A0A085G1G7/1-151 P S Y V C S R P N G L L S D A Q K Q E I A T C I T H S H C E A T G A P P F F V Q V I I E E E G T L K R Y I G A L ----- A T T E - Y I
A0A0D6P319/1-162 P T Y E C R A P A G L D A A R R A R I A A A V T R I H R Q V T G A S A F A Q V I F T E V D A G A W F L G G T ----- P L V G - P H L
A0A0E9NHE3/1-135 P L Y T L Y Y P P G A F T Q W E K P A I A K R I T D V H S S V T G A G A F F V K V I F V P V P A G D I F S G G K ----- L D D Q L V

```

70 80 90 100 110 120 130 140

I FVHGLHREGRSADLKGQLAQRIVDDVSVAAEIDRKH I - WYFGEEMPAQQM - VEYGRFLPQPG - HEGEWFNDLSSDER  
I WQATIRSGRTEKQKEELLRLTQEI ALILGIPNEEV - WYI TEIPGSNM - TEYGRLLMEPG - EEEKWFNSLPEGLR  
A IIRGQIRAGRPQATRAH I LRA I TDVYMAVTDGANAV - VVAVVDIPASWA - ME - - - - -  
LMFVVDIFEGRTEEQRNALIAALSKTGTE TTGIPESDV - RVRLDFPKANMGMAGGISAKAMG - R - - - - -  
PVIVAILIAGRTDEQKRALIAALSETSASVLDAPLQAT - RVMIKDIPNTDF - GIGGQTARALG - R - - - - -  
VMLYGQIRHGRTRDDQKAEIVSQMQASVVEHTGLAPEAV - HVFTTDTASW - MEGGDI LPEPG - EDEWLERHNAKNH  
TSLTLLIRTGRTADYKRGLLTRLWLVLQSATGASDEE I - VLG IHEVPPSQA - MEMGKLMPNVS - ESPADTQ - - - - -  
TLIFCHIRDRSDGAVLSLAKKVSTIWSACTGATEDEV - EVLVTLYPAKYV - VRGGERLPEAP - RV - - - - -  
TVIFCHIRKGRTDGAIERLLKTI SHYAKFTGLGLDEI - EVAAAEYPAVHT - MRNGQLLPEPP - IV - - - - -  
TLIFAHIRAGRSDEAIQSLLKSI SEAWSRI TGDSEDNI - ELAVQQYPAKFT - MRGGRRLLPEPP - IV - - - - -  
TLIMCNI RAGRSDEAKLTLVKKVSIAIFSDYAGVSEDR I - ETGLLEFNPKFI - IRGGQQLPDP - YA - - - - -  
TLIHGNI RVGRSDGAIQKLSKASDI WHDI TQSEDEI - EVAVQEFQKQFV - VRGKPMPEAP - YA - - - - -  
PVAIAIL IAGRTDAQVALIAQLSDAMSA I LDVLPST - RVMIKDIPNTDF - GIGGKTARALG - R - - - - -  
VVFHVMI VGRSEDQKALFTALTKAASSTLGVGENV - RVIVQDVPNTDMGMANGVSAKNTG - R - - - - -  
ALALVRLIAGRDEAKKATLIAALS LAI HASLGAIEQDI - RVVLTDPNTDMGVAGGLTAKAAG - R - - - - -  
ARVDLEI EGRDEAKKALIAALNQAVCASIDISGEDV - EVLRLDVPKTDMGVANGLSAAKAG - R - - - - -  
ALIVDLIAGRTPELKSALISALNQAAACESLGISGQDV - RVVLDHVPKTDMGVANGLSAAAAG - R - - - - -  
ALVTVDMIAGRTDEQKEALIGALNRAVCDSIGIDGTDV - RVMIHVSKANMGVANGISARAAG - R - - - - -  
VALTAAIRSGRSADYKHRLLTSLWTLVKGATGALDEQI - VVGLQEVGPGQA - MEMGRVMP EVD - SNVS - - - - -  
PRMVGLIRDRGTAEVRRALHGIADAWCAVTDGAKADV - AFLHVEVPGANV - LEDGEI LPEAA - DDPVAAHG - - - - -  
LIRGWWRSHPPEDETSQLVAQVAAAATAVTGIPKERV - RVI I IENSPARFA - IEGGRVLP EPG - QERAWLEAH - - - - -  
PVIVAILIAGRTDEQKRALIAALS DAGANVLDAPLQAT - RVI I KDIPNTDF - GIGGQTARALG - R - - - - -  
PIV FALLIAGRTAAKQALIAQIDRACVTVLGA TREPS - RIFIKDIPNTDF - GLAGQTARSLG - R - - - - -  
VLVSGWTRAGHPDAEITRLATEIAAAVTR IAGIPAERV - MVVFDSSPAHYA - VEGGRVLP EPG - HEQAWIAGSG - - - - -  
LLIHGTTRAGRDAEKVRLAKSISTASSEITGVPSERV - LVIIITDPARFA - VEGGRVLP EPG - DEDDWLDEQSN - - - - -  
LLISGWARRGHPQEETRLALELSAAASRI TGIPERRI - LVVIQDSPARSA - VEGGQVLPDPG - QEKEWLSRHEA - - - - -  
LLINGWARRGHPQDSTRLALAEAAASRISGIPQEHV - MVVILDSPARSA - VEAGRVLPDPG - HEAEWLAAGRS - - - - -  
LLINGWRTGHPEAQSSQLVAEIAAAATRV TGVP AERV - LVVIQNSPAHFA - IEGGRVLPAPG - EEEAWLREQKDTD - - - - -  
SVAQFLIAGRSIEQRRLIAGLTVVMAALPGVDSGGV - RII I KDIPNTDF - GIAGQTAQSLGRGIDRSAMAAAAR - - - - -  
ASIVGIRSR - RSTEVKAQLLNDLWSMFKNV TGLSDDQL - WSVTEIPPSNA - MEFGA I IPEVG - REAEWQASLGLTKE  
V FVHGHIRSGRTTDDQVKLLGDI LGSVRQVTGLDSRYV - WIY LSELPSPDM - IEYQVLPQPG - AESAWLQALSEEDR  
L IINGWRTGHSDEQTTALVTQVADAA TRI TGIPAERV - LVIIIGNSPARFA - IEGGRVLPDPG - QELAWLAATTEQSS  
VRFVYRRGRNPARMEQLLQRLYRSVR - VGQPDETVDIEMHVQEADADVW - TLNGVNLPPQGSDEEQRWNTACGVPKQ  
LWIHADTRPGRTEKKTAMIDRMVKEVSTAGGVDESYI - WYVNEI - SEM - AEFMHPFPPG - MEGAFVASLPVEVR  
TFIAGTIRSGRLEQRQLLRELSAAWHELTGQPEEEL - VLSISEQDASAV - MEAGL I FPEAG - AEAWFQONREKLG  
SYLFGAIRHGRDAE TRQTMLEFRSMRSRATGQSEAEF - LVALTEVDPANA - MEAGLVLP EPG - REQEWFENRVRT  
VYVHGLVREGRSIEIKQALMSQMLEEIQIVDI TAEDV - WIY LQDIPATQM - IEFGRFLPAPG - GEAWEKGMTPEKH  
SFLGGQIRHGRSVE TRQAMLKALRDMWQTGQSEAEEL - IVGISEVDPRMV - LEAGFFMPEPG - QEKAWEFEEHARLA  
VWVRADIRSGRTNEQKANI LRRVMRE TSEILGIAEQAV - WYISDIPAQGV - LEFGNVLP EPG - GEEQWLASLPSALR  
V FVHGHIRAGRTPEQKVALLDALSGVVREV LGVPRRTV - WYLVLDLPPADM - IEYGYVLP AAG - EEAEWLASLDDDT  
SGITGSIRAGRTLEVQKLVKDI AASWTSITGQSPKQL - IVGLNEIDSDIT - MEYGLI LPHPG - GEAWFATHADELD  
SFI AAEIRAGRDLDTRQALLRELSGIWTEATGQNEAQL - LVAIKE TPAENA - MEAGLMFPKPG - EEAQWMSSENSDKLA  
VWIHADIRSGRTAQKTDLLEQITSKVADVLELPPEHV - WYVNEIPGENM - TEYKLLPEPG - KEEWFATLPQSLQ  
IWI RADIRSGRTQEQQDQLMR I ADEVSEIAGTSKESV - WYISDIPGSPV - LEFGR I LPPPG - EEDTWFAALSVELQ  
VWI RADIRAGRTVEQKRALLERI TVEVGA I LGLPPEEV - WYVCDIPGSSI - AEHGRVLP EPG - GEDAWFDALPPDLQ  
VWI RADIRAGRTTEETKAE LLRQIVAEVGDITGVVAEHV - WYLVNDVPGPNI - AEYGRVLPNPG - QEDRWFEQLPRELQ  
VFLHGHIRAGRTNVIKKKLIEKLRDSIKKDLNLSKDQV - WYI SELEPSQM - IEYGEI LPKSG - QEKWFNQLPKKLLK  
VTVHGTIRAGRTAEQKSRLLLEDIVGIVAGAANTGRRHV - WYLAELPPAQM - AEYGRVLPQPG - GEAGWLEALPAEDR  
L FVHGHQIRAGRSADQKRDLEALVTLVANATGAEKRSV - WYLVSELPSPDM - VEYKVLPEPG - AESEWLSNMPEPDR  
IHVNGQIRAGRNAEQKRRLLDAIVELVTRAAQAEPRAV - WYIADVPPSQM - VEYGRVLP EPG - EEAANWLQAMPDEDR  
VYVHGHQIRAGRDGETERLVLELMNAVADAAEMPAHCV - QIYVVDVPARQI - AEYQQLLP LPGA - GEAWWWAAIPAE LR  
I FVHGHIRAGRTTEEQKALLADIINSIHGVTGIEKRFL - WYI SE LAPNNM - VEYQVLPQPG - KEAEWLESLSAADR  
AIVHGHIRAGRTPEQKRLLESIVEV IISATALERRYV - WYI SE LPSPDM - VEYGRVLP EPG - AEEDWLKSMDEADR  
VNI VAFLLAGRNDLEKADFMAAINKAAV TSLDVS DTCI - RTMVIDVAPEHMGVQEGLSAKAFR - ARSAS - - - - -  
V FVHGMIRSGRGI DVKQKLMGEAADKVRASAGVGAEDV - WYIQDIPAPQM - IEFGRFLPEPG - AEDAWRKAITPEKQ  
I F I HGHQIRAGRSEEDKRKLEKILLVCS PAAAAPASAL - WYLVLDLPPAQM - AEFGLI LPEPG - QEAQWASLPAADR  
I FVYGHIRSGRAAVDKTRMIRLMADAI RAAANVDS TRAV - WYLVNE LQPRQM - IEFHVLPEPG - DEPAWTEALPDADR  
V FVHGHIRAGRSALRAALVRRRLTDDVAA TARVDR LAV - WYLVSEIPAAAM - VEFHVLPEAG - QEAATEALPAADR  
L FVHGHVRSGRTDHQRMLVERLVRDVAAEASGLP TRA I - WIY LSEIRPSLM - AEFHVLPEPG - EEAWFALPEDDR  
I WVRADIRAGRTKSQLQLMLALKTDIASITDVAEEDI - WIY LNNIEPDNM - LEYGHVLP LPGA - EEKRFWALPAELQ  
IWI RADIRAGRSEEQLKALMLRIVGDVSRISGIPAENV - WINMCM LNP TDM - VEYGRVLP GPG - EEQQWFE TLPQLR  
LFLHGHIRAGRSAGDRMLR I RRLVAA LAEGASLPPRSV - WYVTELP PRAM - AEYHTLP EPG - DEAAWFALPEDQ  
VRI TGVIRAGRDTEARQK I LRGVWEGMQEFLGERKC - - - EMHLEEM LGENV I AENGFMPESGTDEEKRWNGGPM - - -





**Figure A.15 Examples of the curation process used to validate the non-Pro-1 sequences in the TSF.**

As it was surprising that multiple sequences lack an N-terminal proline and are represented globally across the superfamily, it was important to validate this result in detail. Initial bioinformatics identified 1,801 non-eukaryotic sequences without an N-terminal proline. (Eukaryotic sequences were not examined due to the complex nature of their splicing maturation.) Each nucleotide sequence was examined manually and most were removed from the set due to misannotation of the position of the initiating Met or other technical issues (see below). Ultimately, 346 sequences were validated as missing an N-terminal proline.

An example of a properly annotated sequence missing Pro-1 is shown in A.

```

E2PB85      MISVYGLKQTLADRRALIADVIFDCMQMSLGVPKQRHALRFDLLDAENFYPPINRSQDFI
W0R8J4      MISVYGLKHTLAPRRALIAEIIIFSCLEVNLGIPKQRHALRFELDDENFYLPVNRSDNFL
I3DIP6      MITVYGLKKS LAPYRKQIADAIHFHCLHIGLGIPPRKHTLRFVGLEKEDFYLPINRTERFI
I3DR40      MITVFLKSKLAPRREQLAEVIYNSLYLGLDIPKKGHAIRFLCLEKEDFYFPLDRSDDYT
A4N5E7      MITVFLKSKLSRREQLAEVIYNSLHLGLDIPKKGHAIRFLCLEKEDFYFPLDRSDDYT

E2PB85      GGTGGAATGGTTATTAGGAGAAAAAAATGATTAGTGTATATGGATTAACCAACCTTAG
W0R8J4      AGTTCAATGGCTATTGGAGGCAAAAAATGATTAGTGTATACGGATTAACACACACTTG
I3DIP6      TGGCTATTAGCTCGTAAGGAGACATTATGATTACCGTTTATGGTTAAAAAATCACTTG
I3DR40      TTTAACCGCACTTTAAAGGAGAAAAAATGATCACCGTATTCGGACTTAAATCCAAACTCG
A4N5E7      TTTAACCGCACTTTAAAGGAGAAAAAATGATTACTGTATTCGGACTTAAATCCAAACTTT

```

**A. Top.** Multiple sequence alignment of five TSF homologs for which Met-1 is not followed by a proline, which is a deviation from the signature feature of TSF members. However, these sequences clearly align very well. **Bottom.** For all five sequences, their annotated start codons are associated with a Shine-Dalgarno-type sequence, indicating that the annotated start codon is likely the correct one. Indeed, no alternative start codons can be found nearby in which Met would be followed by Pro and where this alternative start codons is also associated with a Shine-Dalgarno-type sequence.

An example of two misannotated sequences missing an N-terminal Pro are shown in B) and the same examples in which the misannotation has been corrected by identification of the true Met1 site is shown in C).

```

A0A0C1MDE4  -----MPIVNIDLIA-RSQDQLKALVQDVTAVTKNTGAPAEHVHVLREMQPNRYGVA
H1LG62      MEEFIMPIVDIHLIA-RSQAQLKGLVEDVTAADVKNKGAPAEHVHVLSEMOKDRYSVG
C0WTL0      -----MPIVNIQLIA-RSQDQLKALVADVTAADVKNKGAPAEHVHVLDEMOKNRYSVG
J1GJS3      -----MPYVNIIRVTREGVSAEQKLALIEGVTDLLEQVLNKKPADTFVVIDEVETDNWVG
F8G1U2      -----MTNEGVS AEHQRQLIEQTTTRMLEQVLGKPPASTFVVEEVP TDNWVG
A0A0P7D6Z0  -----MPYVHIRVTDEGVSAEHRQLIEQTTCLERVLGKPPASTFVVEEVP TDNWVG

A0A0C1MDE4  TAAGAAGAGAAGGGACGAAACTCAATGCAATCGTAAACATCGACTTAAT
H1LG62      GCGTATACTTTTTTAAACTAAAAATGAGGAAATTTATTAATGCAATTGT
C0WTL0      TTAATAAATTTGGAGGAACCTCATCATGCAATCGTAAATATTCAACTTAT
J1GJS3      GGGGTTTGTAACAGGAGAGCCGCTATGCAATACGTTAATATTCGCGTCAC
F8G1U2      AACCATGCCTTATGTCCATATTCGCGTGACCAATGAAGGCGTCAGTGCTGA
A0A0P7D6Z0  GGCATTAACAGAGGAGCAGCAACCATGCAATCGTAAACATCGACTTAAT

```

**B. Top.** Multiple sequence alignment of six protein sequences from the TSF. For two of these sequences, Met-1 (in red) is not followed by Pro. However, their Met-1 is not aligned with Met-1 of the other four sequences, which are their closest homologs in the TSF. This is a clear sign of start codon misannotation. **Bottom.** The start of the DNA sequence of the genes of the six homologs, including the genomic DNA sequence directly upstream of the annotated start codon (in red). Underlined and in bold are the Shine-Dalgarno-type sequences that designate the presence of a ribosomal binding-site. Clearly, the annotated start codons of the two protein sequences for which Met-1 is not followed by Pro, are not associated with a Shine-Dalgarno-type sequence. However, an alternative potential start codon (in green) is seen nearby in the DNA sequence.

A0A0C1MDE4	MP	IVNIDLIA-GRSQDQLKALVQDVTAVTKNTGAPAEHVHVI	REM	QPNRYGVAGVLKS
H1LG62	MP	IVDIHLIA-GRSQAQLKGLVEDVTA	AAVVKNTGAPAEHVHVI	LSEMQKDRYSVGGVLKS
C0WTL0	MP	IVNIQLIA-GRSQDQLKALVADVTA	AAVVKDTGAPAEHVHVI	LDEMQRNRYSVGGVLKS
J1GJS3	MP	YVNIQVIRVTREGVSAEQKLALIEGVTDLLEQV	LNKKPADTFVVIDEVEDN	WGVNRESVS
F8G1U2	MP	YVHIRVTNEGVS	AEHKRQLIEQTT	RMLEQVLGKPPASTFVVEEVP
A0A0P7D6Z0	MP	YVHIRVTNEGVS	AEHKRQLIEQTT	CMLERVLGKPPASTFVVEEVP

A0A0C1MDE4	TAAGAA	<u><b>GAGAAGGGAA</b></u>	CGAAACTCA	<b>ATG</b> CCAATCGTAAACATCGACTTAAT
H1LG62	AAACT	TAAAAAAT	<u><b>GGAGGAA</b></u>	TTTATT <b>ATG</b> CCAATTGTAGATATTCATTTAAT
C0WTL0	TTAAAAA	TTT	<u><b>GGAGGAA</b></u>	CCTCAT <b>ATG</b> CCAATCGTAAATATTCACCTTAT
J1GJS3	GGGGTTT	TGGTAA	<u><b>AGGAGA</b></u>	GCCGCT <b>ATG</b> CCATACGTTAATATTCGCGTCAC
F8G1U2	GGCATT	AAC	<u><b>AGAGGAGCGA</b></u>	GCAACC <b>ATG</b> CCTTATGTCCATATTCGCGTGAC
A0A0P7D6Z0	GGCATT	AAC	<u><b>AGAGGAGCGA</b></u>	GCAACC <b>ATG</b> CCTTATGTCCATATTCGCGTGAC

**C. Top.** Multiple sequence alignment of the six protein sequences of panel B, where the alternative start codon of the two non-Pro-1 sequences is used as the true start codon. Clearly, Met-1 is now followed by Pro. **Bottom.** The aligned new set of DNA sequences. The two alternative start codons are now associated with a Shine-Dalgarno-type sequence, indicating that the alternative start codons are likely the actual start codons of the genes. The fact that Met-1 is followed by a proline, is additional evidence that the start codons are now correctly annotated.

**Figure A.16 PDB codes of structures used in the structure similarity network provided in Figure 2.8.**

1DPT  
1FIM  
1GYX  
1HFO  
1MFI  
1MIF  
1MWW  
1OTF  
1OTG  
1U9D  
1UIZ  
2AAG  
2OP8  
2ORM  
2OS5  
2WKB  
2WKF  
2X4K  
2XCZ  
3ABF  
3B64  
3C6V  
3E6Q  
3EJ3, chain B (*cis*-CaaD)  
3EJ3, chain C (*cis*-CaaD)  
3FWT  
3GAD  
3KER  
3M20  
3MB2, chain B (*cis*-CaaD)  
3MB2, chain C (*cis*-CaaD)  
3MF8  
3N4G (CgX)  
3RY0  
3T5S  
4DH4  
4FAZ  
4FDX  
4JCU  
4JJ9  
4LHP  
4LKB  
4M1A  
4OTA (4-OT)

4U5P  
5UIF (Linker 1)  
5UNQ (Linker 2)  
6BLM (Fused 4-OT)

Notes: Structures used in **Figure 2.8** and in the MSA for **Figure 2.5** differ for 4-OT and *cis*-CaaD. For 4-OT, PDB 1BJP is used in the MSA for **Figure 2.5** and PDB 4OTA is used in **Figure 2.8**. For *cis*-CaaD, PDB 2FLZ is used in the MSA for **Figure 2.5** and PDBs 3EJ3 and 3MB2 are used in **Figure 2.8**. Two chains each from *cis*-CaaD structures 3EJ3 and 3MB2 were used in **Figure 2.8** as these proteins are organized as part of physiological heterohexamers in which the B and C chains are different.

## **Appendix B. Supplemental Information for Chapter 3**

### **B.1 Sequence alignments and HMMs for level-1 subgroups of the TSF**

The HMMS and sequence alignments and HMMs generated are listed along with their github URLs.

#### **B.1.1 5-(carboxymethyl)-2-hydroxymuconate isomerase**

sequence alignment: <https://github.com/babbittlab/tsf/blob/master/alignments/chmi.clustal>

HMM: <https://github.com/babbittlab/tsf/blob/master/hmms/chmi.hmm>

#### **B.1.2 *cis*-3-chloroacrylic acid dehalogenase**

sequence alignment: <https://github.com/babbittlab/tsf/blob/master/alignments/cis-CaaD.clustal>

HMM: <https://github.com/babbittlab/tsf/blob/master/hmms/cis-CaaD.hmm>

#### **B.1.3 Macrophage migration inhibitory factor**

sequence alignment: <https://github.com/babbittlab/tsf/blob/master/alignments/mif.clustal>

HMM: <https://github.com/babbittlab/tsf/blob/master/hmms/mif.hmm>

#### **B.1.4 Malonate semialdehyde decarboxylase**

sequence alignment: <https://github.com/babbittlab/tsf/blob/master/alignments/msad.clustal>

HMM: <https://github.com/babbittlab/tsf/blob/master/hmms/msad.hmm>

## **B.2 Sequence alignments and HMMs for level-2 subgroups of the 4-oxalocrotonate tautomerase subgroup**

### **B.2.1 Group 1**

sequence alignment: <https://github.com/babbittlab/tsf/blob/master/alignments/4OT-group1.clustal>

HMM: <https://github.com/babbittlab/tsf/blob/master/hmms/4OT-group1.hmm>

### **B.2.2 Group 2**

sequence alignment: <https://github.com/babbittlab/tsf/blob/master/alignments/4OT-group2.clustal>

HMM: <https://github.com/babbittlab/tsf/blob/master/hmms/4OT-group2.hmm>

### **B.2.3 Group 3**

sequence alignment: <https://github.com/babbittlab/tsf/blob/master/alignments/4OT-group3.clustal>

HMM: <https://github.com/babbittlab/tsf/blob/master/hmms/4OT-group3.hmm>

### **B.2.4 Group 4**

sequence alignment: <https://github.com/babbittlab/tsf/blob/master/alignments/4OT-group4.clustal>

HMM: <https://github.com/babbittlab/tsf/blob/master/hmms/4OT-group4.hmm>

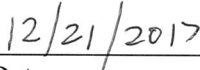
**Publishing Agreement**

*It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.*

***Please sign the following statement:***

*I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.*

  
\_\_\_\_\_  
Author Signature

  
\_\_\_\_\_  
Date