Published on Dec 01, 2022 by Arcadia Science **DOI**: 10.57844/arcadia-ynth-kh70

Defining actin: Combining sequence, structure, and functional analysis to propose useful boundaries

The process of deciding whether a candidate actin homolog represents a “true” actin is tricky. We propose clear and data-driven criteria to define actin that highlight the functional importance of this protein while accounting for phylogenetic diversity.

Contributors (A-Z)

Prachee Avasthi, Brae M. Bigge, Feridun Mert Celebi, Megan L. Hochstrasser, Taylor Reiter

Version 7 · May 20, 2025

Purpose

To learn about a protein’s function and regulation across a broad range of species, you must define which of the many potentially related proteins you’re going to count as homologs and where the line between true homologs and other proteins exists. Then, understanding the proteins that exist at this boundary can help identify novel functions and regulation, as well as insights into how the protein family evolved. Determining

whether a protein fits within a particular protein family requires characterizing the sequence, structure, and importantly, the function of that protein.

We've outlined a series of well-defined and testable criteria for determining whether a candidate actin is a "true" actin as opposed to an actin-related protein or an actin-like protein. Using these criteria, we created a pipeline to computationally analyze candidate actins. We ran almost 50,000 candidate actins through this pipeline and, among other things, found that global sequence conservation and functional analysis showed a distinct cluster of true actins.

These criteria and the pipeline we developed to analyze them might be useful for anyone studying "fringe" actins. We would love feedback on whether you think these criteria are sufficient, if there are other criteria we should include, and what might make this pipeline more useful for your own work.

- This pub is part of the **project**, "Annotating proteins based on critical functions." Visit the project narrative for more background and context.
- All associated **code** is available in this GitHub repository. If you want to run your own actin candidate through the pipeline, you can use the Binder here and follow the instructions here.
- The **data outputs** from our actin identification pipeline are available here.

Motivation

Let's say you're mining genomes for homologs of your favorite protein, and you see a protein that looks promising. Ideally, you would try to understand how closely the sequence, structure, and function of the candidate protein match your main protein of interest using computational and experimental tools. You look at the sequence similarity using a program like BLAST to compare the sequence to a known protein **[1] [2]**. With the recent advances in AlphaFold and structural comparison, you might also run the predicted structure of your favorite protein through a comparison search, and that can tell you about the structural similarity of your protein to others **[3][4]**. But how

do you know when those results mean that a protein is similar enough to be relevant or considered a homolog?

Another important characteristic when wading through possible homologs is protein behavior, or function. Typically, we explore function in an experimental setting, where we first identify an intriguing potential homolog based on sequence or structure, and then investigate it in an experimental system. For example, if you think that a particular protein might be involved in cell division, you might mutate that protein in your cells and see if it affects division, or tag that protein and see where it localizes when cells divide. This can tell you a lot of information about your protein, but it is generally pretty low-throughput and requires a lot of time and effort.

As part of our project to functionally annotate proteins, we wanted to overcome this issue for specific protein families. We wanted to choose a family of proteins to use as our first use case that is responsible for a wide range of cellular functions, well-studied enough that we know or can predict how it performs its most basic functions, and diverse enough that there's plenty of room for discovery.

Our use case: Actin

Actin is a cytoskeletal protein that is required for a long list of cellular functions that are essential for life, and it is sorted into these diverse functions through its interaction with actin-binding proteins (this list is not exhaustive and is periodically updated and refined). Because it's important for so many functions, actin is generally well-conserved and present throughout the tree of life. However, most of what we know about actin comes from cells that represent a relatively small sliver of the tree of life, mostly Opisthokonts (amoebae, fungi, and animals) with very highly conserved actins. Because of this, our rules about what makes an actin an actin might be incomplete. This means that we are missing out on important data about actin, its functions in the cell, and what determines which of these many functions actin will perform in a given species or at a given time. We are also potentially missing out on how we might be able to re-engineer cellular functions based on the regulation of actin and what is possible in a wide range of organisms. Therefore, it is important to look at actins that lie right on the boundary between "true" actins and actin-like or actin-related proteins, making this an interesting use case for our functional annotation pipeline.

Unsurprisingly, the first thing we did was a BLAST search against the NCBI non-redundant database using human ß-actin. This gave us a list of about 50,000 related proteins, but we realized there are no clear rules about how similar an actin has to be to our model actins to be considered a true actin. We did structure searches and found a similar problem. None of this really told us if the proteins we were looking at were similar enough to be considered actins but different enough to potentially provide new insights.

This is complicated by the presence of actin-related proteins and actin-like proteins. Actin-related proteins, or ARPs, are a class of proteins found across cell types that are highly similar to conventional actin, but that have different cellular functions, different abilities to polymerize, and are generally only found in cells that express a separate, primary actin. ARP1, for example, is part of the dynactin complex that forms with dynein. ARP1 is able to form short filaments within the complex, but is unable to form longer independent filaments. ARP2 and ARP3 are part of the Arp2/3 complex, which nucleates new branched actin filaments. They serve as the first two subunits of the newly forming actin filament. Other ARPs are important for chromatin remodeling and mitochondrial dynamics **[5]**.

Actin-like proteins are present in cells that already express a primary actin, or in non-eukaryotic cells. An example of #1 is the novel actin-like protein 1 (NAP1) in *Chlamydomonas reinhardtii*, *Volvox carteri*, and other closely related algae that encode a primary actin. *Chlamydomonas* NAP1 is roughly 60% identical to mammalian actin, while the primary actin is closer to 90% **[6][7]**. Non-eukaryotic examples are actin-like proteins in archaea, including Crenactin and Lokiactin, and bacteria, including MreB and ParM.

There are no clear rules for when a candidate homolog should be considered an actin, an actin-related protein, or an actin-like protein. Here, we work towards defining a “true” actin by creating a set of clear, easily testable, and quantifiable criteria that can be used to functionally annotate these proteins. Beyond actin, we hope that this general workflow and the idea of using quantitative measures of similarity across sequence, structure, and function to define protein families will be broadly useful.

The proposed criteria and our actin identification pipeline

The quantifiable criteria we propose to define actin are as follows (click to jump to our analysis for each criterion):

1. Sequence conservation
2. Structural conservation
3. Functional conservation

In narrowing down our initial list, we considered a few important things. First, the importance of each criterion to the overall function of the protein helped us determine which criteria really mattered. We were more likely to consider criteria that are very important for actin function, like polymerizability, than those that do not necessarily influence the function of the protein, like phylogeny. Additionally, we selected criteria that we could easily determine for our candidate actins. We can determine most of the criteria using computational tools or simple experiments.

Using the three criteria above and each step described below, we created a streamlined and efficient pipeline that tells us the likelihood that a protein of interest is a true actin. While other tools allowed us to look at global sequence identity or structural identity independently, this pipeline considers sequence and structural identity together as well as important functional properties and their conservation (Figure 1).

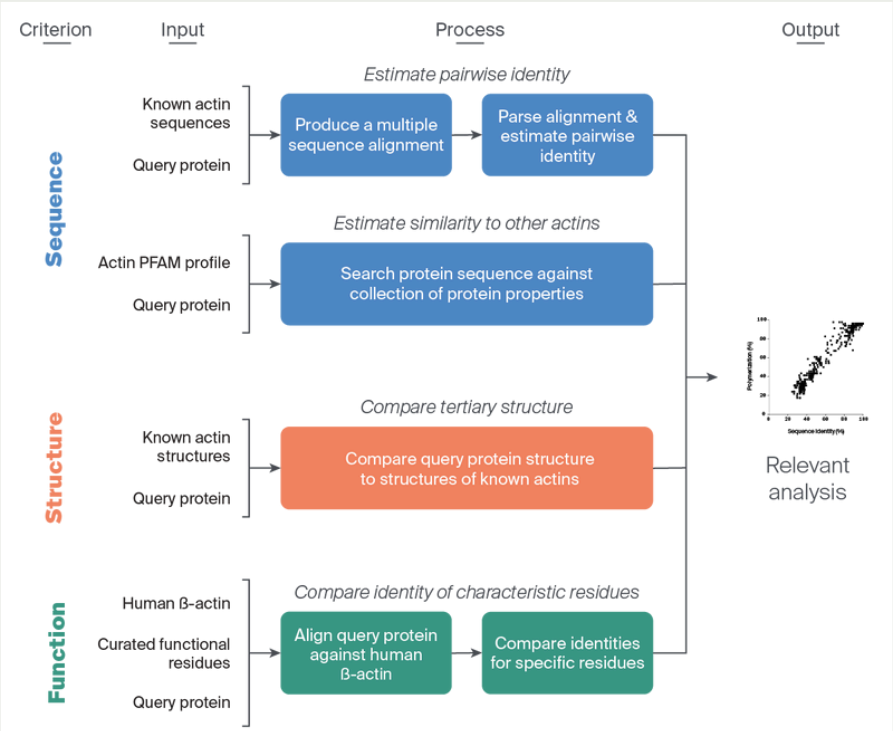

Figure 1

The actin prediction pipeline.

To investigate whether our proteins of interest are true actins, we analyzed a group of actins based on their sequence, structure, and function.

Computational method for actin identification

Briefly, we used the pipeline to perform a global sequence analysis by comparing query proteins to a multiple sequence alignment containing frequently studied actins that we know polymerize using MAFFT **[8][9]**. We also used the actin PFAM profile to determine if the proteins of interest were members of the actin family using the hmmer3 package **[10][11]**. Next, we determined structural conservation by comparing structural models of query proteins that were determined using AlphaFold to a known actin structure using the Foldseek program **[3][4][12]**. Finally, we looked at specific actin functions by aligning query proteins to human ß-actin labeled with specific residues that are known to be important in either polymerization of the protein or its ATPase function again using MAFFT **[8][9]**. More information on this pipeline to

investigate the “actin-ness” of a particular protein of interest can be found in subsequent sections and on GitHub.

> All **code** generated and used for the pub is available in this GitHub repository (DOI: 10.5281/zenodo.7622712), including all of the processes summarized in Figure 1.

Applying the pipeline

While we use this specific pipeline to look at actins, the idea behind this pipeline is broadly applicable to other proteins. Coupling sequence and structure analysis together in a fast and efficient pipeline and adding in a functional component can help better define various families of proteins and can help researchers determine whether or how their proteins of interest fit within those families.

To test this pipeline, we performed analyses of all of the candidate actins that came up when we did a BLAST search of human ß-actin (UniProt ID: P60709), arbitrarily limiting the output to the first 50,000 sequences **[1]**. Of these 50,000 initial BLAST matches, 2,363 failed to download from NCBI with eutils (error invalid uid), returning empty FASTA files. So, we analyzed 47,634 candidate actins. We outline our key findings in the next section.

Findings

Sequence conservation shows clustering of “true” actins and other proteins

Generally, a protein’s global sequence identity to a known protein is used to determine its divergence or similarity to other proteins. The amino acid sequence, or the primary structure, helps determine how the protein will assemble into its secondary structure. It is also important for interacting with other proteins, with other monomers in the case of actin, and with other molecules in the cell, like ions and small molecules. Thus,

looking at the global sequence similarity can be a useful metric. However, this metric alone ignores other ways in which proteins can be similar resulting in likely misses of proteins that may be related. It could also give rise to spurious relationships between proteins that have actin-like sequences but do not function like actin.

Most actins consist of about 375 amino acid residues. Humans have six actins, which, compared to each other, are at least 93% identical (Figure 2, A) **[13]**. Most of the differences in these sequences appear at the extreme N-terminus, where these differences cause differential regulation due to their post-translational modifications. On the other end of the spectrum, the most divergent eukaryotic actin currently characterized belongs to the single-celled parasite, *Giardia*, coming in at roughly 58% sequence identity compared to human actin **[14]**. Between *Giardia* and humans lies a wide spectrum of actins that could be potential goldmines for better understanding actin biology, and this doesn't even consider the vast array of actin family proteins that exist outside Eukarya. This again underscores the need to clearly define actin-related proteins, actin-like proteins, and true actins.

We approached this issue in two ways. First, we used MAFFT to create a multiple sequence alignment that consists of extensively studied, known actins that function as we would expect, including human actins, yeast actins, and several other conventional actins (Figure 2, A) **[8][9]**. We then aligned each of our actins of interest (the top 47,634 results from running human ß-actin through BLAST) to this multiple sequence alignment and calculated an average pairwise identity. This tells us how conserved our actins of interest are to a set of known, well-studied conventional actins. Next, we looked at the conservation of our actins of interest in relation to the actin family of proteins by comparing our sequences to the actin PFAM profile using hmmer3 **[10][11]**. This primarily tells us whether a given candidate actin fits into the broader family of actin proteins.

Because we initially identified proteins based on sequence similarity, we found that all of the query proteins we analyzed do align well with the actin PFAM profile and therefore do fit into the actin family. Next, we determined the average global sequence identity of each query protein compared to the multiple sequence alignment in panel A (Figure 2, B–C). Average global sequence identity of the query proteins ranged from about 25% to nearly 100% (Figure 2, B–C). The data appears to be multimodal with about 4–6 peaks and a noticeable transition in the data between about 60–70% (Figure 2, C).

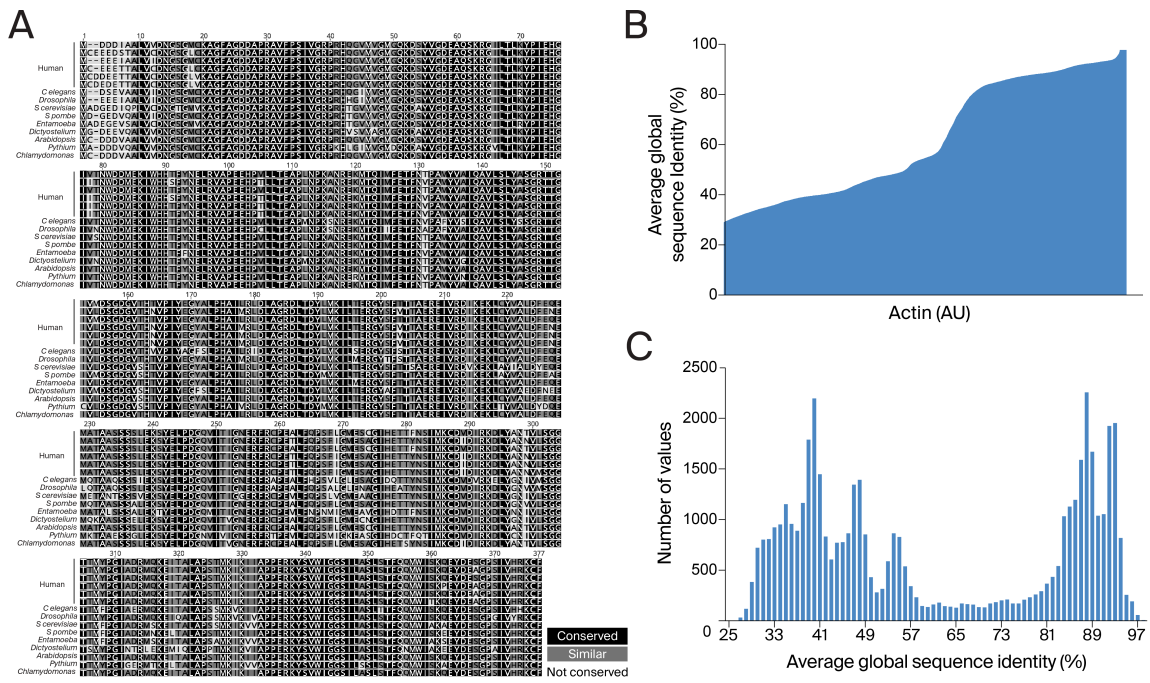

Figure 2

Global sequence conservation shows clustering of actins into bona fide actins and other proteins.

(A) The multiple sequence alignment used for this step of the analysis containing known actins that are well-studied and that have been shown to polymerize normally.

(B) The average global sequence identity of all of the actins shown compared to the multiple sequence alignment shown in A. AU refers to arbitrary units; the x-axis in this case is each of the individual actins.

(C) Frequency distribution of the data graphed in C showing clear clustering of query proteins into true actins and other proteins.

Structural conservation does not align well with sequence conservation, highlighting the need for multiple analyses

Structural conservation adds another dimension to assessing protein relatedness, and the advancements made with AlphaFold, the AlphaFold Protein Structure Database, and programs like Foldseek make this relatively simple to implement into our pipeline **[3][4][12]**.

The actin fold is composed of four subdomains, termed subdomains 1–4, which have varying degrees of importance in actin's polymerization, ATPase function, and binding ability (Figure 3) **[15]**. Between subdomains 2 and 4 lies the nucleotide-binding cleft, where actin binds and hydrolyzes ATP. Between subdomains 1 and 3 lies the target-binding cleft (TBC), where a large number of actin-binding proteins bind, including profilin, gelsolin, WH2 domain proteins like WASP, and FH2 domain proteins like formins. Additionally, this region is highly important for the association of actin monomers during polymerization.

This "actin fold" is critical because in order to perform its cellular functions, actin must be able to polymerize, perform its ATPase functions, and interact with a large range of actin-binding proteins. Due to its functional importance across organisms, the overall structure of actin is well-conserved. A similar fold can be found in actins from humans to the distant eukaryote *Giardia*, but also in archaea and bacteria (Figure 3, A–D). This characteristic actin fold is also shared by non-actin proteins, including some actin-related proteins, sugar kinases, hexokinase, and Hsp70 proteins **[16]**. This highlights the importance of considering multiple criteria in determining whether a protein is indeed an actin.

To include structural analysis in our pipeline, we obtained structural models for a set of our candidate actins using the AlphaFold Database **[3][4]**. We compared these proteins to the experimentally determined structure of rabbit muscle actin (PDB: 1J6Z) using Foldseek **[12]**. Foldseek works by turning 3D protein structures into 3Di sequences based on geometric information regarding residues and their nearest neighboring residues **[12]**. It then aligns these sequences and produces a list of scores (E-values) associated with each alignment. The Foldseek E-values are similar to BLAST E-values in that they show the statistical likelihood that a protein is similar as opposed to random **[1]**. For complete information on how Foldseek calculates E-values

see their publication **[12]**. We use them here to tell us how structurally similar our query proteins are to our reference protein.

Of the 50,000 candidate actins we identified via BLAST, we analyzed structures of those that were on UniProt and had structures determined by AlphaFold (17,036). After compiling these scores, we found a distribution of structural conservation (Figure 3, E). We compared the structural conservation to the sequence conservation, and interestingly, did not see any obvious pattern in structural conservation and sequence conservation (Figure 3, F). Specifically, there is a density in the high sequence conservation space that has a huge range of structural divergence, which we did not expect.

We suspected this might be due to partially sequenced proteins, often denoted as "partial proteins." To test this, we used the length of the protein sequences to color the graph in Figure 3, F, based on the fact that most actins are around 375 amino acids in length **[17]**. We confirmed that this unusual distribution is due to the presence of partial proteins — proteins less than 300 amino acids long account for the bulk of the confusing data. To avoid this in the future, we will check the quality of our input data and consider adding filtering steps.

This kind of large-scale structural analysis could also be useful for understanding which aspects of the actin fold are conserved elsewhere, like in sugar kinases and heat shock proteins, and making predictions to annotate functions in non-actin proteins.

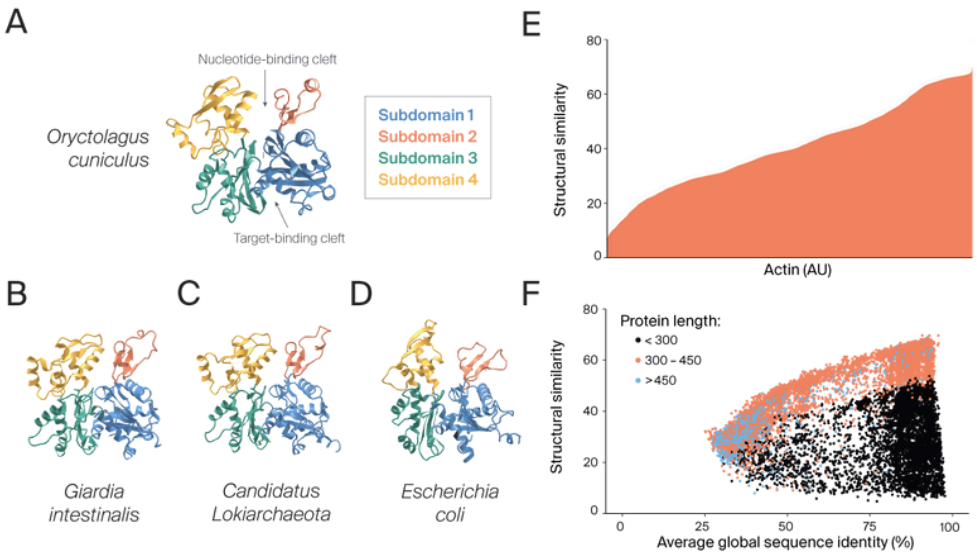

Figure 3

Actin structure is broadly conserved, but differences between structural conservation and sequence conservation highlight the importance of doing multiple analyses.

(A) Cytoplasmic actin from rabbit (PDB: 1J6Z) with subdomains represented in different colors. Of note are the nucleotide-binding cleft where ATP is typically bound and the target-binding cleft which is important for the binding of several known actin-binding proteins.

(B) AlphaFolded actin model of the most divergent eukaryotic actin currently known, the actin from *Giardia* (AlphaFold Database: AF-P51775-F1). Subdomains of the protein colored as in A. The overall structure is quite similar.

(C) AlphaFolded actin model of the Archaeon *Candidatus Lokiarchaeota* (AlphaFold Database: AF-A0A532TFF0-F1). Subdomains of the protein colored as in A. The overall structure is quite similar.

(D) AlphaFolded actin model of the bacterial actin-like protein, MreB (AlphaFold Database: AF-P0A9X4-F1). Subdomains of the protein are colored as in A. Despite the evolutionary distance and lack of sequence identity, the overall structure is still quite similar.

(E) Structural similarity represented by –1*log transformed E-values obtained from FoldSeek for each candidate actin.

(F) Bivariate analysis of –1*log-transformed E-values, showing structural similarity and the global percent identity. Colors of the dots correspond to the length of the proteins in amino acids.

Actin's polymerizing function is less conserved than its ATP-binding ability

After the determination of sequence and structural conservation, we are often forced to turn to the bench to determine functional conservation. Usually this would be done by looking at how the protein functions in the cell or *in vitro.* However, we hoped to address this issue computationally and in a relatively high-throughput manner by identifying the residues important for specific protein functions and looking at their conservation. This is possible because proteins are multifunctional and different domains of proteins have different functions that might set that functional criteria. While the multiple sequence alignments, Hidden Markov Models, and FoldSeek used previously will work for basically all proteins as long as there is a good reference protein, the residue-specific functional annotations are typically the rate-limiting step. Because we already know a lot about actin, we know that polymerization and ATPase activity are key functions to probe through this approach.

Polymerization

Polymerization is actin's ability to transition between a monomeric state and a filamentous state. During polymerization, monomers interact with each other to form a polymeric filament (Figure 4, A). This important characteristic is found in all actins and many actin-like proteins, but actin-related proteins are unable to form long, stable polymers. This suggests that a candidate protein's ability to polymerize is a highly important characteristic in deeming it a "true" actin.

Typically, researchers use pyrene assembly assays and total internal reflectance fluorescence (TIRF) microscopy to experimentally test actin polymerization. But is there a way to computationally predict whether a putative actin will likely polymerize?

Actin filaments form a right-handed helix composed of two chains of actin monomers **[15]**. So, polymerization of actin involves two types of interactions between monomers: longitudinal (long-pitch) contacts between monomers within a chain and lateral (short-pitch) contacts between monomers in adjacent chains (Figure 4, A). Based on cryo-EM structures of actin filaments, researchers have found the residues involved in each of those types of contacts (Figure 4, B) **[18]**. Using this information, we created a computational program that looks for conservation of those particular residues as a way to determine if polymerization of a potential actin is likely.

We ran our candidate actins through this program and got results for 20,095 candidate actins that had non-gapped alignments to our reference human ß-actin. In a subsequent version, we plan to adjust the pipeline to include alignments with gaps as well. Here, we found that there is a broad range of conservation of these residues for both lateral contacts and longitudinal contacts (Figure 4, C–D). Putting these contacts together, we saw a clear transition between a group of query actins that are more conserved and a group that is less conserved (Figure 4, E–F). We also looked at polymerizability compared to both sequence identity (Figure 4, G) and structural conservation (Figure 4, H). We found that in both cases, there is a correlation between the percentage of conserved residues involved in polymerization and each of the other two criteria.

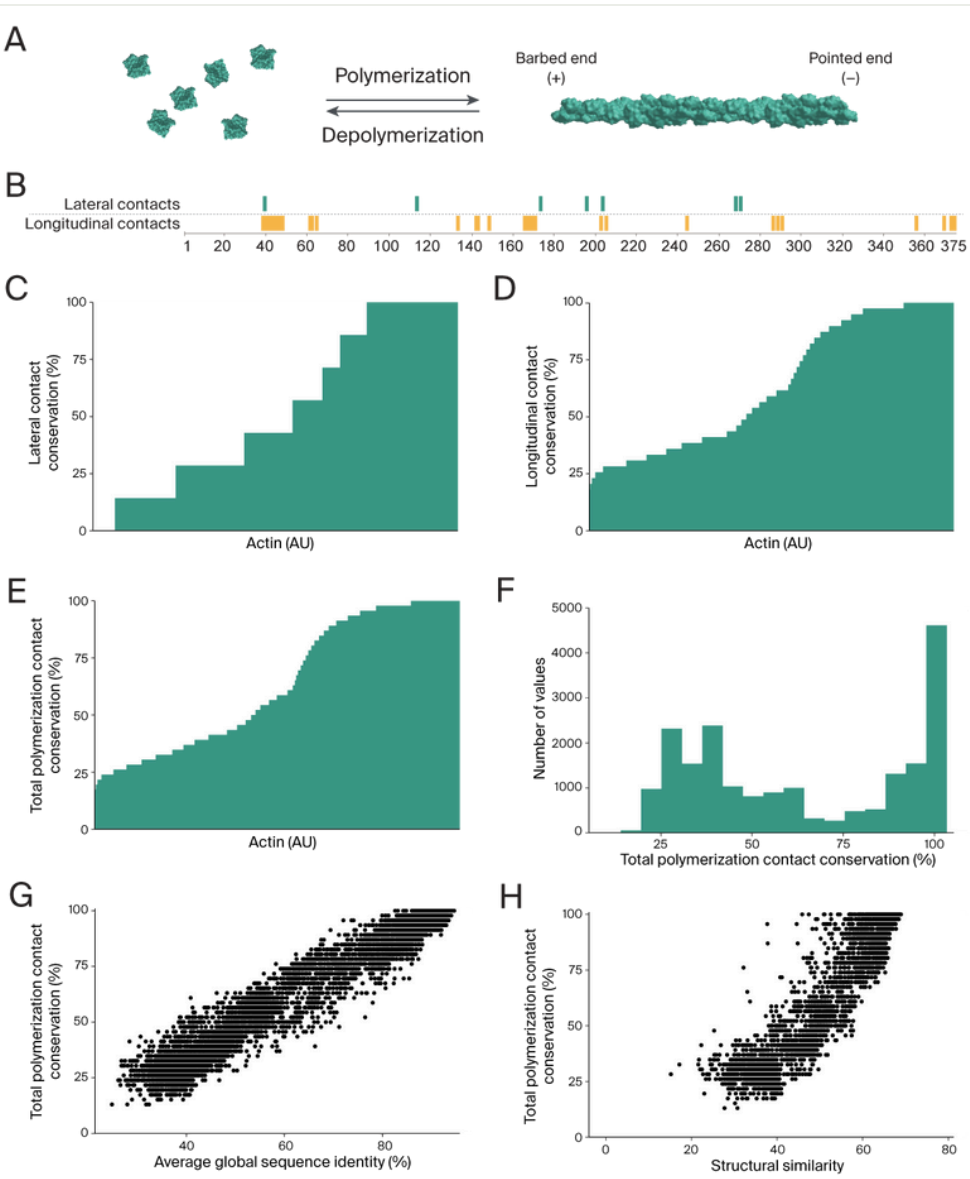

Figure 4

Analysis of the conservation of the residues involved in specific actin functions reveal clustering of proteins into actins and other proteins.

(A) Actin monomers (PDB: 1J6Z) polymerize into polar filaments composed of two helical actin chains (PDB: 3G37).

(B) Annotation of all actin residues involved in lateral and longitudinal contacts between monomers within filaments.

(C) Percentage of lateral contacts conserved throughout the query actins.

(D) Percentage of longitudinal contacts conserved throughout the query actins.

(E) Total polymerization contacts (lateral and longitudinal) conserved throughout the query actins.

(F) Frequency distribution of the total conserved polymerization contacts showing a cluster of well-conserved actins and a cluster of less-conserved actins.

(G) Bivariate analysis comparing the polymerizability to the global sequence identity showing correlation between the two, with some outliers.

(H) Bivariate analysis comparing the polymerizability to the structural similarity, represented by -1*log-transformed E-values, showing correlation between the two, with some outliers.

ATPase activity

Important for polymerization and depolymerization, actin functions as an ATPase, an enzyme that hydrolyzes ATP. Typically, monomeric actin bound to ATP joins the end of a growing actin filament. As the filament ages, the ATP is hydrolyzed to ADP and inorganic phosphate (Pi) (Figure 5, A) **[19]**. Then, once the inorganic phosphate is released, ADP-bound actin can be released from the filament as monomeric actin. The ADP in monomeric actin is swapped out for ATP so the monomers can rejoin actin filaments once again.

ATPase function can be determined with biochemical assays. However, similar to polymerization, the region of actin that binds nucleotides is known based on crystal structures and cryo-EM structures of actin with several different bound nucleotides (ATP, ADP, ADP + Pi) (Figure 5, B) **[18]**. Using these residues, we created a program that looks for conservation of the nucleotide-binding site to use as a readout of possible ATPase function.

We aligned each of query sequences to that of human ß-actin and annotated the regions that have been found to be involved in ATP binding and therefore ATPase function. We analyzed 32,680 of our original candidate actins that had non-gapped

alignments to our reference human ß-actin to our reference human ß-actin and found that overall, the ability to bind ATP seems to be more conserved than the ability to polymerize (Figure 5, C–F). Even some proteins with relatively low percent global identity or structural similarity still seem likely to bind ATP based on the conservation of the residues involved (Figure 5, E–F). However, there are cases where these residues are not well-conserved. These might be interesting targets for better understanding how actin functions as an ATPase and how those functions evolved.

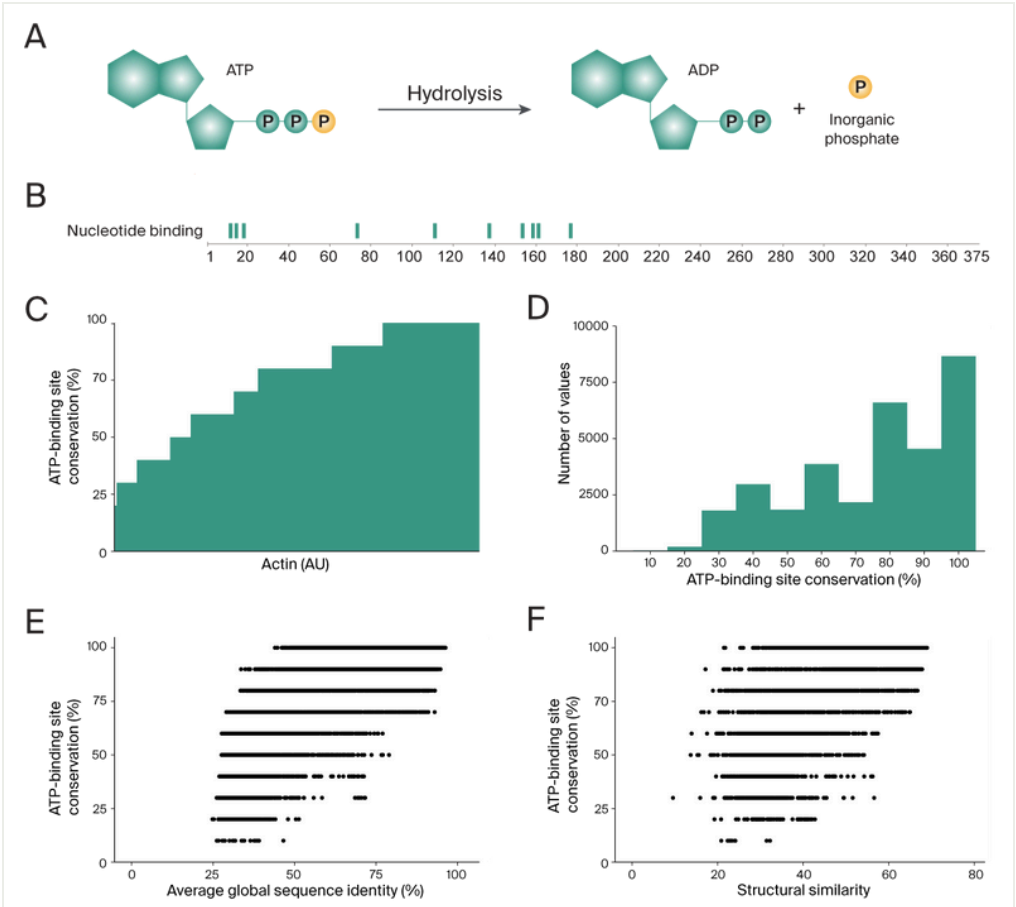

Figure 5

Analysis of the conservation of the residues involved in specific actin functions reveal clustering of proteins into actins and other proteins.

(A) Actin binds ATP and hydrolyzes it into ADP and inorganic phosphate.

(B) Annotation of actin residues that are important for nucleotide binding and hydrolysis.

(C) The conservation of the residues predicted to bind ATP in our query actins.

(D) Frequency distribution of the data in C.

(E) Bivariate analysis comparing the potential for ATPase activity to the global sequence identity.

(F) Bivariate analysis comparing the potential for ATPase activity to the structural similarity, represented by -1*log transformed E-values.

Key takeaways and conclusions

Key takeaways

- Analysis of global sequence identity shows clustering of potentially true actins and other proteins (Figure 2).
- Structure analysis and sequence analysis are not well-correlated, highlighting the need for evaluating multiple criteria instead of relying on one (Figure 3).
- The conservation of residues involved in polymerization is well-correlated with global sequence identity and shows similar clustering of true actins and other proteins (Figure 4).
- Generally the residues that bind ATP are well-conserved (Figure 5).

Conclusions

Based on the key takeaways summarized above, we returned to our original problem of defining the line between true actins and other proteins so that we can identify divergent proteins that exist at this border. We wondered whether we could use the patterns we observed to determine which proteins are actins as opposed to actin-like proteins or actin-related proteins. So we investigated how existing annotations fit with the data presented here. We extracted gene names for the proteins that were listed on UniProt and parsed these into one of three categories: actin (this includes any type of actin and any isoforms labeled specifically as “actin”), actin-like proteins (this includes both proteins termed actin-like proteins and actin family proteins), and actin-related proteins (this is any and all actin-related proteins or ARPs). We then visualized our results with these categories mapped on the graphs.

Looking at the percent identity alone, we found that our large peak between about 80–100% is primarily composed of “actins” (Figure 6, A). The majority of proteins between

57% and 80% also seem to be “actins” (black), while proteins with lower percent identities are primarily annotated as “actin-like” (purple) or “actin-related” proteins (yellow) (Figure 6, A). However, we also found that the designation of “actin-like” does not necessarily mean that a protein has a lower sequence identity, underscoring the need for a multi-dimensional tool like this to determine where candidates fit within the broader family of actins and actin-like proteins.

We also mapped these existing annotations onto our analysis of structural similarity compared to global sequence identity (Figure 6, B). Here, while the percent identity of “actin-related” proteins (yellow) is much lower, the structural similarity of these same proteins is around average (Figure 6, B). Meanwhile, “actins” and “actin-like proteins” have a broad range of structural conservation, which was unexpected (Figure 6, B). This could be due to the high structural conservation throughout the entire actin family. Because most actin family proteins are highly structurally conserved, perhaps we are just not able to see clear patterns among them.

Finally, we mapped these existing annotations onto our analyses of functional similarity compared against global sequence identity (Figure 6, C–D). For the polymerization, it is clear that “actin-related” proteins seem to be clustered in a region of low global sequence identity and also low conservation of the polymerization contact sites, while most “actins” seem to have high conservation of both (Figure 6, C). “Actin-like” proteins, however, are spaced out across the distribution, suggesting that this particular annotation is not as meaningful as we might think (Figure 6, C). All the proteins seem to have relatively well-conserved ATP-binding sites. The clustering of “actins,” “actin-like” proteins, and “actin-related” proteins is less clear, likely because the ATPase function of actin extends into related proteins, even “actin-like” proteins and “actin-related” proteins (Figure 6, D).

Together, these data demonstrate the importance of considering multiple criteria when deciding whether a protein fits within a protein family. Each of our criteria — sequence, structure, and protein function — yield slightly different results and distributions, but considering the full picture provides more insight into when a candidate could be considered a true actin.

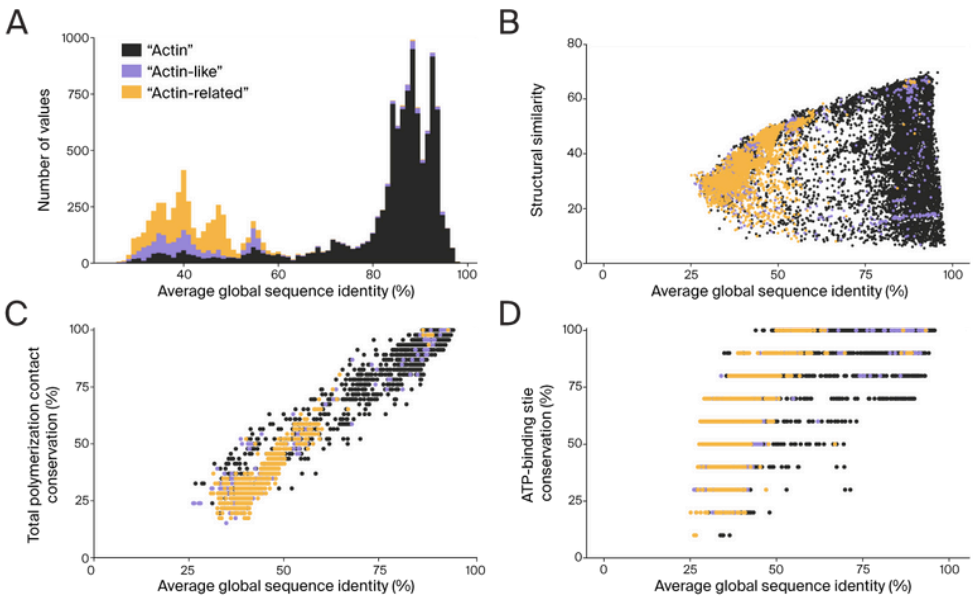

Figure 6

Labeling of candidate actins based on existing annotations shows interesting patterns.

(A) Frequency distribution of the candidate actins that are present on UniProt and have relevant annotations. Colors correspond to existing annotations.

(B) Bivariate analysis of structural similarity as determined by the log-transformed E-values and the global sequence identity for all candidates with relevant annotations available. Colors correspond to pre-existing annotations.

(C) Bivariate analysis of the conservation of the residues important for polymerization and the global sequence identity for all candidate actins with annotations available. Colors correspond to existing annotations.

(D) Bivariate analysis of the conservation of the residues involved in ATP binding and the global sequence identity for all candidate actins with annotations available. Colors correspond to existing annotations.

What do you think?

The purpose of this work is to create a useful and clear metric to decide when a protein is an actin as opposed to an actin-related protein, an actin-like protein, or some other protein entirely. In order to do this, we built a pipeline that we think could be broadly applicable to other actin researchers, but that could also be adapted by protein biologists anywhere who are interested in a specific protein family. We hope this tool can serve as a framework for the creation of similar tools for different protein families.

This list of specific and testable criteria is designed as a starting point for a larger discussion about how we determine the definitions of cytoskeletal proteins and other proteins. How might criteria such as these be applied to cytoskeletal proteins, or even non-cytoskeletal proteins, that are expressed throughout the tree of life?

We hope that this list of criteria will be useful for researchers studying actin and cytoskeletal proteins. Do you feel that these criteria could be useful to determine whether a protein is an actin or not? Are there places we should be more or less specific? Is there anything we didn't include that you feel we should include, or anything we included that you don't find relevant?

How could we expand the current scope of our pipeline to make it useful to your own research? Are there ways that we could take this beyond studying actin, and if so, what protein families should we expand to?

Finally, if you want to run your own actin candidate through the pipeline, you can use the Binder here and follow the instructions here, and let us know how it goes!

We would love any feedback or thoughts you'd like to contribute.

---

1 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990). Basic local alignment search tool. https://doi.org/10.1016/s0022-2836(05)80360-2

2 STATES DJ, GISH W. (1994). QGB: Combined Use of Sequence Similarity and Codon Bias for Coding Region Identification. https://doi.org/10.1089/cmb.1994.1.39

3 Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A, Žídek A, Green T, Tunyasuvunakool K, Petersen S, Jumper J, Clancy E, Green R, Vora A, Lutfi M, Figurnov M, Cowie A, Hobbs N, Kohli P, Kleywegt G, Birney E, Hassabis D, Velankar S. (2021). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. https://doi.org/10.1093/nar/gkab1061

4 Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. (2021). Highly accurate protein structure prediction with AlphaFold. https://doi.org/10.1038/s41586-021-03819-2

5 Goodson HV, Hawse WF. (2002). Molecular evolution of the actin family. https://doi.org/10.1242/jcs.115.13.2619

6 Onishi M, Pringle JR, Cross FR. (2015). Evidence That an Unconventional Actin Can Provide Essential F-Actin Function and That a Surveillance System Monitors F-Actin Integrity in*Chlamydomonas*. https://doi.org/10.1534/genetics.115.184663

7 Kato-Minoura T, Uryu S, Hirono M, Kamiya R. (1998). Highly Divergent Actin Expressed in aChlamydomonasMutant Lacking the Conventional Actin Gene. https://doi.org/10.1006/bbrc.1998.9373

8 Katoh K. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. https://doi.org/10.1093/nar/gkf436

9 Katoh K, Standley DM. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. https://doi.org/10.1093/molbev/mst010

10 Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, Finn RD, Bateman A. (2020). Pfam: The protein families database in 2021. https://doi.org/10.1093/nar/gkaa913

11 http://hmmer.org/documentation.html

12 van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, Söding J, Steinegger M. (2022). Fast and accurate protein structure search with Foldseek.

https://doi.org/10.1101/2022.02.07.479398

13 Perrin BJ, Ervasti JM. (2010). The actin gene family: Function follows isoform. https://doi.org/10.1002/cm.20475

14 Paredez AR, Assaf ZJ, Sept D, Timofejeva L, Dawson SC, Wang C-JR, Cande WZ. (2011). An actin cytoskeleton with evolutionarily conserved functions in the absence of canonical actin-binding proteins. https://doi.org/10.1073/pnas.1018593108

15 Dominguez R, Holmes KC. (2011). Actin Structure and Function. https://doi.org/10.1146/annurev-biophys-042910-155359

16 Bork P, Sander C, Valencia A. (1992). An ATPase domain common to prokaryotic cell cycle proteins, sugar kinases, actin, and hsp70 heat shock proteins. https://doi.org/10.1073/pnas.89.16.7290

17 Elzinga M, Collins JH, Kuehl WM, Adelstein RS. (1973). Complete Amino-Acid Sequence of Actin of Rabbit Skeletal Muscle. https://doi.org/10.1073/pnas.70.9.2687

18 Chou SZ, Pollard TD. (2019). Mechanism of actin polymerization revealed by cryo-EM structures of actin filaments with three different bound nucleotides. https://doi.org/10.1073/pnas.1807028115

19 Kanematsu Y, Narita A, Oda T, Koike R, Ota M, Takano Y, Moritsugu K, Fujiwara I, Tanaka K, Komatsu H, Nagae T, Watanabe N, Iwasa M, Maéda Y, Takeda S. (2022). Structures and mechanisms of actin ATP hydrolysis. https://doi.org/10.1073/pnas.2122641119