A Framework for Inferring Protein Location as a Function of Condition

Shannon Quinn

CMU-CB-10-101
April 29, 2010

Lane Center for Computational Biology
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Robert F. Murphy (Chair)
Aarti Singh
Hagit Shatkay (U. of Delaware)

Submitted in partial fulfillment of the requirements for the degree of Master of Science.

Keywords: subcellular localization, conditions, clustering, location patterns, similarity

Abstract

The Waldo framework offers one method for determining subcellular protein location patterns. The framework operates by gathering data from many different protein databases. Waldo builds a model from proteins with observed location patterns, clustering them by their locations under specific conditions. This creates clusters whose labels are effectively consensus location patterns. These consensus patterns serve as starting points for proteins whose location patterns under the conditions of interest are unknown. Under the assumption that similar proteins localize similarly, the unobserved proteins are compared against those that are clustered, identifying a cluster whose constituent members have structure and sequence closest to that of the protein of interest. By associating a known protein of closest match to the unknown protein, the location pattern of the known protein provides a point estimate for the unobserved location pattern. Using the Z-score and percent identity resulting from the homology comparison, a confidence can then be placed on the point estimate of the location pattern.

1 Introduction

The problem of protein subcellular localization is multifaceted: proteins must be identified (e.g. by GFP-tagged fluorescence microscopy), monitored over discrete or continuous time points, and observed in specific cellular compartments. Furthermore, proteins can be subjected to variable conditions (e.g. a cancerous cell [1]) which may have the effect of altering their locations within the cell. Particularly in the Murphy lab, the primary methods of determining protein locations under differing conditions have been through automated image analysis [2] and text mining [3].

By virtue of proteins appearing in multiple locations under differing conditions, we use the phrase "location pattern" to describe the distribution over cellular compartments in which proteins are found. The goal of the Murphy lab has been to locate and identify all proteins under all conditions, such as the filaman protein illustrated in Figure 1. The Waldo framework, an approach which relies heavily on data that has been gathered in previous proteomics research, is but one of many approaches to reach this goal.

An important aspect to protein localization is the biochemistry involved, including the structure and sequence of the protein. The approach employed by Waldo makes use of the assumption that similar proteins will express similar localization behavior. This approach depends on a measure of both protein similarity and localization similarity. In addressing the former, there are already numerous methods in place to quantitatively discern a measure of similarity between proteins, from double dynamic programming [4] to structural distance matrix alignment [5, 6]. BLAST [7], a local alignment tool, is yet another option which locally aligns protein sequences to determine peptide similarity.

2 The Waldo Framework

2.1 Phase I: Data Gathering and Waldo Architecture

Waldo is written in Python, and is a standalone application that can be deployed on any system running Python 2.6 or greater (requires packages Amara and SQLAlchemy). The first step in using Waldo to infer protein location is the aggregation of a significant amount of data published with previous research. At this time, the data sources used by Waldo consist of five separate protein databases. The names and number of distinct proteins available from each data

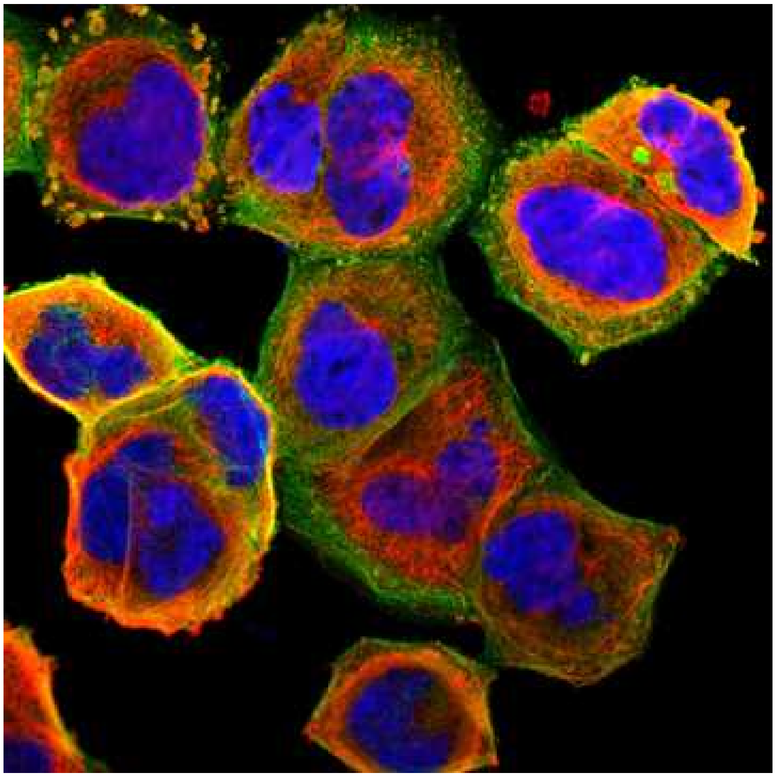


Fig. 1. A-431 cell line of HPA protein with associated Antibody identifier 2925. This is an alpha filamen A, seen here localizing as the green fluorescent antibody in the cytoplasm and cytoskeletal actin filaments. (http://www.proteinatlas.org/show_image.php?image_id=200004293&channel=_blue_red_green)

source are listed in Table 1. As Waldo initializes itself, this data is downloaded from each source and normalized as it is saved, enabling comparison of the data from different research databases.

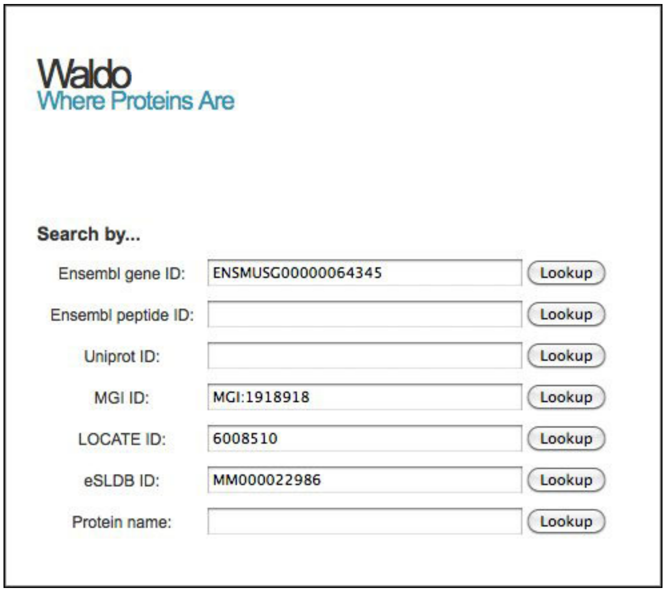
Database	URL	Entries
LOCATE[8]	http://locate.imb.uq.edu.au/	122,765
MGI[9]	http://www.informatics.jax.org/	34,162
eSLDB[10]	http://gpcr.biocomp.unibo.it/esldb/	80,220
UniProt[11]	http://www.uniprot.org/	516,081
HPA[12]	http://www.proteinatlas.org/	10,100

Table 1. The total number of unique entries (with the exception of HPA: the immunofluorescence data has three entries per antibody, one for each cell line) from each data source that is tracked by Waldo is indicated in the third column.

Data normalization involves two separate processes. First, Waldo internally converts each database-specific identifier (a Uniprot accession, for example) to a global namespace: for the current iteration, each database-specific identifier was translated to a matching Ensembl Gene identifier. This conversion allows for the comparison of proteins regardless of their data source of origin. Most databases already provided this identifier within the annotations of the data. For those which only supplied a database-specific identifier, or which provided only an Ensembl Peptide (such as eSLDB), the Synergizer service [13] was used to translate the identifiers between namespaces.

The second aspect of data normalization involves structuring the data properly. A specific database schema (for storing the data locally) and an accompanying web-based interface together provide a means for querying and visualizing the stored information in such a way that allows side-by-side comparison of information such as subcellular protein location, GO terms, accompanying citations, and other annotations. Some data from each source (such as UniProt comments not related to protein location) was not needed in the most recent iteration, so it was not captured in the normalization process.

As mentioned, Waldo can be used as a proxy for each protein database it captures, providing a web front-end for users (shown in Figure 2) to query by any identifier with which Waldo is familiar: any assimilated data source, in addition to any Ensembl identifier. A command-line batch interface is also provided for users interested only in the location patterns for proteins from any of the assimilated databases.



Search by...

Ensembl gene ID: ENSMUSG00000064345
Ensembl peptide ID:
Uniprot ID:
MGI ID: MGI:1918918
LOCATE ID: 6008510
eSLDB ID: MM000022986
Protein name:

Fig. 2. Web-based interface of the latest Waldo iteration. This will be changed soon to a single query field which accepts all identifier types, as the ability to dynamically recognize a specific identifier is already implemented in the latest version of Waldo.

2.2 Phase II: Building an Inference Model from Labeled Data

The majority of the data captured by Waldo does not have condition information associated with it. As such, no assumptions are made regarding the condition of the recorded protein. Only HPA data provides this information in the form of discrete cell lines. These different cell lines, in which each protein has an observed location pattern, are used in the next step of the framework as training data to build the predictive model. This model, in combination with tools to measure protein similarity, are used to predict the location patterns under these three conditions, or cell lines, of unobserved proteins. This process can be generalized to any conditions for which there is training data available; however, these conditions must be made explicit to the model, as learning them in an unsupervised way is beyond the scope of Waldo.

The HPA data set (a few entries of which are shown in Table 2) consists of HPA-specific identifiers associated with a cell line and ensuing location pattern. The representation of the location pattern is an 18-dimension binary vector, where each dimension is indicative of a different cellular compartment of interest (e.g.

Nucleus, ER, Golgi). The bit in each of the vector's elements indicates whether or not the protein was observed in that specific compartment.

Antibody	Cell line	Nucleus	Cytoplasm	Mitochondria	ER	Golgi	...
0609	A-431	0	0	0	1	0	...
4799	U-2 OS	1	1	0	0	0	...
3536	U-251MG	0	0	1	0	0	...

Table 2. Partial location pattern vectors for three different HPA antibodies. For each antibody, there are three separate cell lines, though for brevity only one cell line of each is shown. If the binary vector value is "on" at a specific position, this indicates the protein was observed at the corresponding location.

HPA constitutes the entirety of labeled data for training, exactly 10,100 observations over all three cell lines. Since each protein is observed three times, once for each condition, this equates to 2,969 distinct proteins, each with three location profiles, or vectors. As stated, the cell line attribute is synonymous with a specific condition on the protein, simulating the effect of a drug or mutation to observe the change, if any, in subcellular protein localization.

To construct a predictive model from this information, the data is preprocessed to remove any proteins for which all three location profiles consisted exclusively of zeros, indicating a protein which was unobserved. After preprocessing, the number of distinct and fully observed proteins decreased to 2,770. This formed the primary data set for building the model.

In the next step, these data points were clustered. Each of the three location profiles for every protein were concatenated to generate a single 54-dimensional binary vector, describing the protein's location across all three conditions. Since the attributes for clustering are binary, the Hamming distance metric was used (as it satisfies the Triangle Inequality) to calculate distances between protein location profiles and generate groups, making explicit a certain edit distance between clusters of proteins with similar location patterns across all three conditions. The Bayesian Information Criterion (BIC) was used to determine the ideal K number of clusters, which ranged from 2 up to 100. The BIC equation followed in the form of the Schwartz Criterion, as

$$BIC(M_j) = \hat{l}_j(D) - \frac{p_j}{2} \log R \qquad (1)$$

where $\hat{l}_j(D)$ is the log-likelihood of the data from the jth model, and p_j is the number of parameters (or clusters) in model M_j [14]. The log-likelihood function can be written as a sum of likelihoods over each of the centroids. Therefore, the likelihood of each centroid, given the subset of points D_n from the total points D that has μ_n as their centroid, is given by the following equation:

$$\hat{l}(D_n) = -\frac{|D_n|}{2}\log(2\pi) - \frac{|D_n|M}{2}\log(\hat{\sigma}^2) - \frac{|D_n| - K}{2} + |D_n|(\log|D_n| - \log|D|) \tag{2}$$

where $|D_n|$ is the number of points assigned to the nth cluster, $|D|$ is the total number of points, $\hat{\sigma}^2$ is the maximum likelihood estimate for the variance in the nth cluster under the assumption of normally distributed data, and M is the dimensionality of the data (in this case, $M = 54$). This produces a BIC curve which must be maximized.

As indicated in Figure 3, the BIC for the given K initially increases before reaching a point at which it merely begins oscillating. While the global maximum of this function with respect to the number of clusters is found at $K = 61$, the function remains qualitatively steady across a large interval of K, posing a question of tradeoffs in terms of model complexity and goodness of fit; a lower value of K, while not conferring the largest BIC, would create a less complex and possibly more generalizable model of clusterings. Therefore, in order to further illuminate the effects of increasing K on the BIC, the variance of the BIC was plotted against K as two separate functions, the results of which are shown in Figure 4. One curve shows the amount of variance in the BIC plot up to the specific K, while the other shows the amount of variance in the BIC plot after the specific value of K.

By determining the smallest value for K at which the difference between the two curves is greatest, that value of K represents the simplest and most generalizable model for which the maximum amount of overall variance is captured, theoretically leading to a model which is most effective at capturing relationships between clusters. This point occurs at $K = 46$, which is the global maximum of the blue curve, representing the amount of variance already captured by the specific K. This value also correlates with a local minimum of the second curve, which is very close to the curve's global minimum.

Once the clusterings of proteins based on location pattern is built, it can be queried. Within the parameters of the assumption that similar proteins have similar location patterns, an unobserved query protein is presented to the model in an attempt to determine its location profile based on sequence and structure

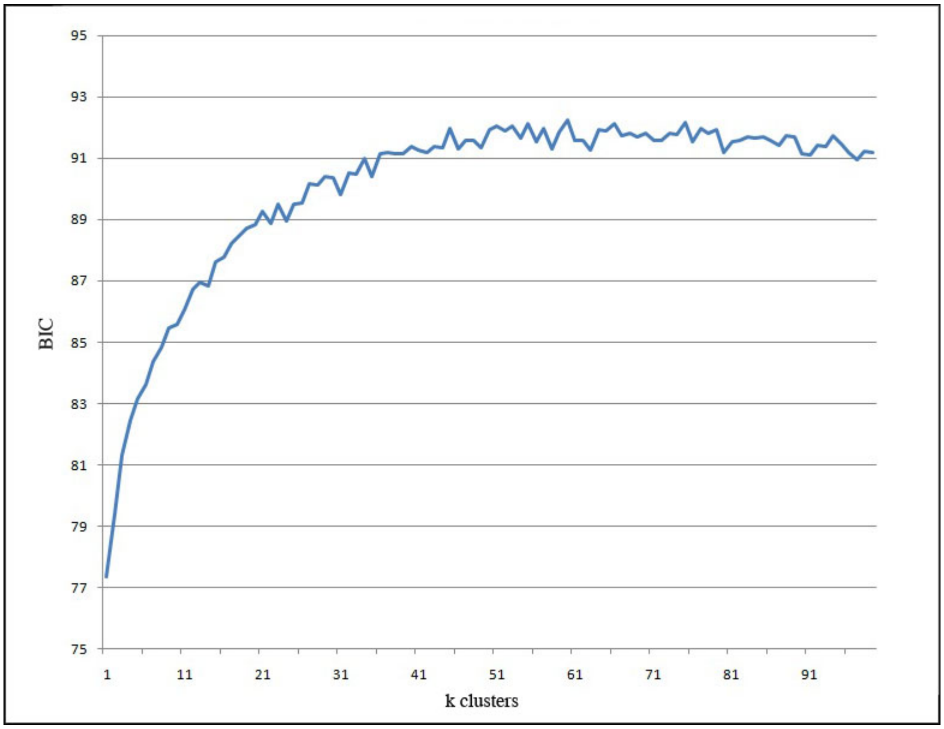


Fig. 3. Plot of the calculated BIC vs 2,770 data points given a number K of clusters. As the data within each cluster is assumed to follow a normal distribution, the likelihood in the BIC equation is the sum of squares [14].

homology. The query protein is compared to representative proteins from each cluster using the DALI pairwise distance matrix alignment [5, 15]. In comparing the PDB structures of the two proteins, a Z-score and percent identity alignment is generated. From these metrics, a best-fit cluster is determined for the query protein. The highest Z-score and the protein which yielded it are then used to construct a confidence assignment. The observed protein's location profile is used as a point estimate for the query protein's location profile, with a confidence level that is proportional to the Z-score.

A second batch of trials was also run, using BLAST as the similarity metric instead of DALI. In this case, the alignments producing the lowest E-value were used.

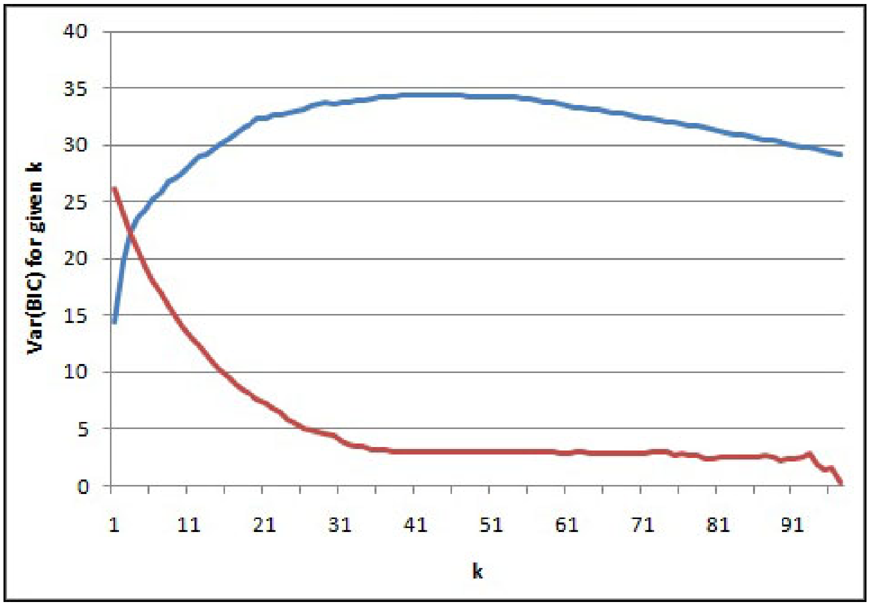


Fig. 4. The red line plots $Var(\text{BIC})$ for all $k \geq K$, whereas the blue line shows the variance for all $k \leq K$. The point at which the difference of these two functions reaches its maximum indicates the K at which the maximum $Var(\text{BIC})$ has already occurred, and the variance that remains is minimal. This provides a way of selecting the smallest K that still captures the majority of the BIC variance, and thereby the majority of the clustering information.

3 Discussion

To use the DALI alignment tool, proteins required an associated PDB identifier. Out of the 2,770 HPA proteins, 809 contained PDB information. Therefore, in using 10-fold cross-validation to test the model, each test fold consisted exclusively of proteins drawn from the 809 with PDB identifiers. 46 clusters were constructed for each iteration, and for each iteration the proteins of the test set were compared to representatives of each of the 46 clusters. When a closest match (highest Z-score) was found, the query protein was tentatively assigned the location profile of the closest matching protein with a confidence proportional to the percent identity of the two peptides.

In order to test the accuracy of the clustering and homology mechanism for inferring location, the percent identity of the query-observed combinations were plotted against the Hamming distance of the observed protein's location profile from that of the query protein in the test fold. The results are shown in Figure 5.

Type	Antibody	Partial Location Profile	Z-score	% Id	H-Dist
Query	3303	0 1 0 0 0 1 1 0 0 0 1 0	25	0.41	2
Match	2831	0 0 1 0 0 1 1 0 0 1 1 1	–	–	–
Query	4177	0 0 1 0 1 0 0 1 0 0 0 1	58	0.83	7
Match	3282	1 1 1 1 1 1 0 1 1 1 1 1	–	–	–

Table 3. Shows partial location profiles of two query peptides and the observed matching proteins (as determined by PDB comparison yielding the largest Z-score and percent identity). Match proteins indicate peptides on which the model was trained; Query proteins are "unobserved" testing peptides to gauge model accuracy.

As is qualitatively apparent in Figure 5, there is not a statistically significant correlation between PDB structure alignment and the Hamming distance of the location profiles. Had the initial assumption of similar proteins localizing similarly been apparent in the data, the scatter plot would have shown a correlation roughly from the top left of the graph sloping to the bottom right. Table 3 illustrates part of the problem: while there are certainly protein pairings which exhibit behavior consistent with the initial assumption, many others – even with relatively high Z-scores and percent identities - localize very differently under the same conditions. The plot shows similar discrepancies: while a qualitative argument could be made for the number of proteins with a relatively low Hamming distance as percent identity increases, the results are nevertheless not statistically significant due primarily to a large amount of noise and an orthogonal lack of data, as only 226 points appear in Figure 5, even though each fold constituted 81 test points (one fold with 80) for a total of 809 data points over all 10 folds.

While the initial assumption is theoretically well-grounded in broad proteomics research, the approach taken here was unable to show any significant correlation of protein similarity with similarity of subcellular location profile. Table 4 reflects a portion of the confusion matrix for all the test folds. Overall accuracy across all 226 points, according to the full confusion matrix, was a mere 3.5%.

The results of the BLAST iterations were very similar. Like the DALI approach, a conversion between identifier namespaces was required, this time to Uniprot accession numbers. The E-scores, or likelihood of seeing a particular hit at random, were used to gauge the quality of the matching and to choose the best. As seen with DALI in Table 3, high-quality E-values did not correlate particularly well with point estimate location patterns. With both DALI and BLAST, consensus location profiles were constructed from a logical disjunction of each constituent location profile, creating a vector which essentially indicated all possible locations to which proteins in the cluster would localize under all conditions. Using these consensus profiles, the Hamming distance was again calculated between the query protein and both the consensus of the cluster to which DALI /

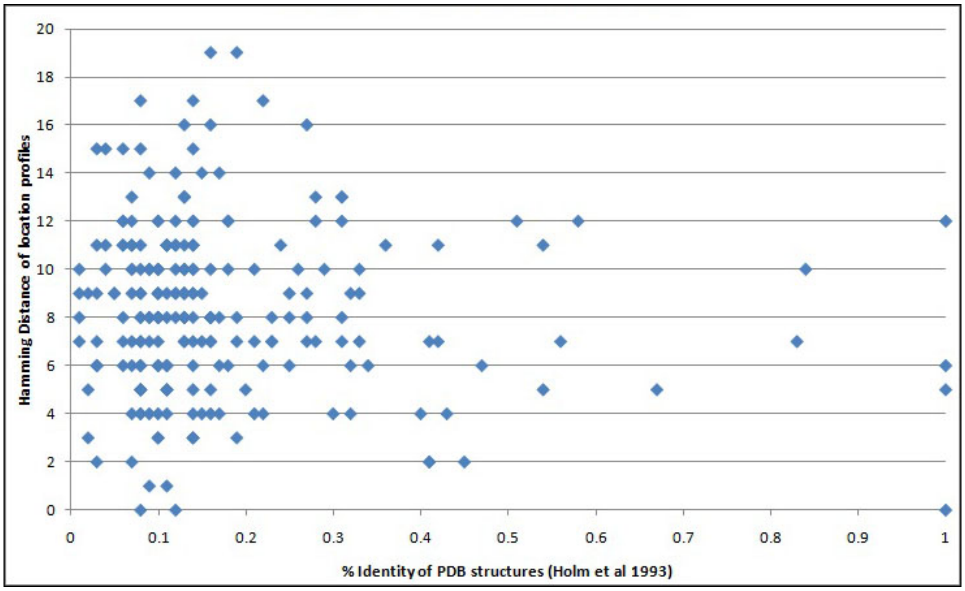


Fig. 5. Plot of hamming distance between query-match proteins vs percent identity as indicated by DALI. Each point represents a best-match pairing of a query protein with a corresponding observed protein using DALI as the criterion, and is plotted to show the relationship of the matching's percent identity (proxy for Z-score) with the Hamming distance (in units of distance of the bit vectors) of the two peptides' location profiles.

BLAST indicated the closest match, and to all other cluster consensus profiles to determine which was the true closest cluster. Results are shown in Figure 6.

4 Conclusion

Presented here is a method for comparing proteins and discerning localization profiles under conditions of interest. This framework relies on the core assumption that proteins of similar structure and sequence will behave similarly, or have similar location profiles. Waldo provides the means for downloading this data, normalizing it, allowing the user to query it, and forms the basis for inferring the locations of proteins for which we do not have location data under predefined conditions.

0	1	0	0	0	0	0	0	0	0	1	0	0	1	0
0	**1**	0	0	0	0	0	0	0	0	0	0	0	1	0
0	0	**0**	0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	**0**	0	1	0	0	0	0	0	0	0	0	0
0	1	0	0	**0**	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	**0**	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0
0	0	0	0	0	0	1	1	**0**	0	0	0	0	0	2
0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	1
0	0	0	0	0	2	0	0	0	0	**0**	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	1
0	1	0	0	0	0	0	0	0	1	0	0	0	**0**	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**

Table 4. A portion of the full confusion matrix, showing the first 15 clusters. Diagonal entries, indicating where proteins that belonged to a certain cluster were assigned to that cluster as part of a test fold, are bolded.

Waldo has its shortcomings, the most egregious of which is a lack of comprehensive data. Much of this problem can be attributed to limitations in the tools used. The DALI database for comparing PDB structures, while indexing all 809 of the available PDB data points, does not contain pairwise comparison data for every possible pairing of the 809 proteins, leaving more than half of each test fold unpaired – only 226 of the original 809 proteins could be paired for comparison – thereby creating results highly susceptible to noise. Furthermore, the 809 proteins with PDB data are but a subset of the available data from HPA: a total of 2770 proteins, nearly 70% for which the Harvard Synergizer could not identify any corresponding PDB data. This contributes to incredible data sparsity, given its 54-dimensional location profiles. The methods presented here also implicitly assume that all location profiles are equally likely, when this is almost assuredly not the case: a qualitative observation will reveal that the majority of location profiles have very few ”on” bits. Taking the specific properties of the location profiles account when building the model may yield better initial clustering.

Alternatively, it is also possible that the format of the location profiles is simply inappropriate for this sort of by-proxy inference. While the location profiles yield very sparse data, their binary nature makes constructing location distributions difficult, particularly without a much larger amount of data.

Nevertheless, there are numerous ways by which the methods described here could be improved to show more definitively whether the initial assumptions are correct. One possibility would be examine several different methods of quan-

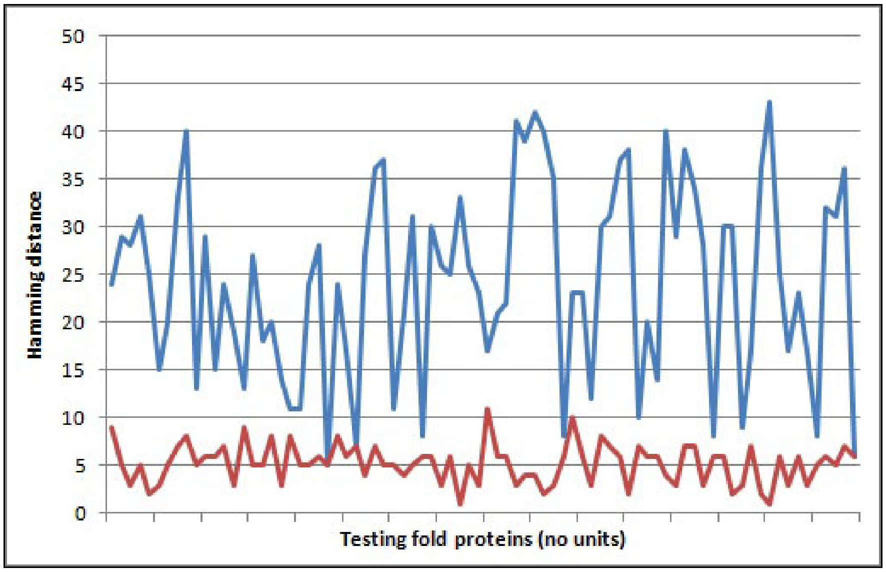


Fig. 6. The blue line plots Hamming distance between query-match proteins, while the red line plots Hamming distance of the query proteins and their true closest match in terms of location profile. Each point on the horizontal axis represents a different query protein. This shows that BLAST / DALI did not perform particularly well as a means for predicting protein location by protein similarity score, as in almost every single case the nearest cluster was not the predicted one, or even close to the predicted one.

tifying protein homology for determining matches: rather than a distance matrix alignment, double dynamic programming methods or hierarchical alignment methods could be explored and compared, ideally allowing more of the original HPA data to be used in training and testing. BLAST would be an ideal tool for performing local alignments between two proteins.

Another method for improvement would entail building two separate clustering models. One model would consist of the Hamming distance K-means clusterings as illustrated here, while the other would be a clustering of the same proteins on homology distance. From these two clustering models, overlap of clusters and constituent members could gauge directly whether similar proteins (as defined by the parameters of the homology clustering) group closely with proteins whose localization profiles are also similar.

Yet another method involves leaving clustering behind and estimating location profiles directly. Since the location profiles are binary, the system could be set up as a series of boolean expressions, and a SAT solver used to correlate location profiles with conditions, or expression solutions. In order to determine the relationships between conditions and the ensuing location profiles, a linear system could be established consisting of matrices $\mathbf{O}_i$ whose rows consist of proteins under a specific condition i, and their subsequent location profiles as the columns. These observed matrices would be a linear function of some unobserved "perfect" protein localization profile $\mathbf{U}$, multiplied by a location-to-location coefficient matrix $\mathbf{C}$, producing the observed mappings. Solving for the unobserved patterns would allow a significant level of generalization for introducing new conditions into the system.

It is important to emphasize that Waldo serves as a starting point, in which any of the above directions could potentially be implemented.

References

1. Glory, E., Newberg, J., Murphy, R.: Automated comparison of protein subcellular location patterns between images of normal and cancerous tissues. In: Proceedings of the 2008 IEEE International Symposium on Biomedical Imaging. (2008) 304–307
2. Newberg, J., Murphy, R.F.: A framework for the automated analysis of subcellular patterns in human protein atlas images. Journal of Proteome Research **7**(6) (2008) 23002308
3. Shatkay, H., Höglund, A., Brady, S., Blum, T., Dönnes, P., Kohlbacher, O.: Sherloc: High-accuracy prediction of protein subcellular localization by integrating text and protein sequence data. Bioinformatics **23**(11) (2007) 1410–1417
4. Flores, S., Echols, N., Milburn, D., Hespenheide, B., Keating, K., Lu, J., Wells, S., Yu, E., Thorpe, M., Gerstein, M.: The database of macromolecular motions. Nucleic Acids Research **34** (2006) D296–301
5. Holm, L., Sander, C.: Protein-structure comparison by alignment of distance matrices. Journal of Molecular Biology **233** (1993) 123–138
6. Koehl, P.: Protein structure similarities. Current Opinion in Structural Biology **11** (2001) 348–353
7. Altschul, S., Gish, W., Miller, W., Myers, E., Lipman, D.: Basic local alignment search tool. Journal of Molecular Biology **215**(3) (1990) 403–410
8. Sprenger, J., Fink, J.L., Karunaratne, S., Hanson, K., Hamilton, N.A., Teasdale, R.D.: Locate: a mammalian protein subcellular localization database. Nucleic Acids Research **36** (2008) D230–D233
9. Bult, C.J., Eppig, J.T., Kadin, J.A., Richardson, J.E., Blake, J.A., the Mouse Genome Database Group: The mouse genome database (mgd): mouse biology and model systems. Nucleic Acids Research **36** (2008) D724–D728
10. Pierleoni, A., Martelli, P.L., Fariselli, P., Casadio, R.: esldb: eukaryotic subcellular localization database. Nucleic Acids Research **00** (2006) D1–D5
11. The UniProt Consortium: The universal protein resource (uniprot) in 2010. Nucleic Acids Research **38** (2010) D142–D148
12. Berglund, L., Bjorling, E., Oksvold, P., Fagerberg, L., Asplund, A., Szigyarto, C.A.K., Persson, A., Ottosson, J., Wernerus, H., Nilsson, P., Lundberg, E., Sivertsson, A., Navani, S., Wester, K., Kampf, C., Hober, S., Ponten, F., Uhlen, M.: A

genecentric human protein atlas for expression profiles based on antibodies. Molecular & Cellular Proteomics **7**(10) (2008) 2019–2027

13. Berriz, G., Roth, F.: The synergizer service for translating gene, protein, and other biological identifiers. Bioinformatics **24**(19) (2008) 2272–2273
14. Pelleg, D., Moore, A.: X-means: Extending k-means with efficient estimation of the number of clusters. In: Proceedings of the 17th International Conf. on Machine Learning. (2000) 727–734
15. L, H., S, K., P, R., A, S.: Searching protein structure databases with dalilite v.3. Bioinformatics **24** (2008) 2780–2781