University of Nebraska at Omaha

DigitalCommons@UNO

Interdisciplinary Informatics Faculty Proceedings & Presentations

School of Interdisciplinary Informatics

2010

An intelligent data-centric approach toward identification of conserved motifs in protein sequences

Kathryn Dempsey Cooper
University of Nebraska at Omaha, kdempsey@unomaha.edu

Benjamin Currall
Creighton University

Richard Hallworth
Creighton University

Hesham Ali
University of Nebraska at Omaha, hali@unomaha.edu

Follow this and additional works at: https://digitalcommons.unomaha.edu/interdiscipinformaticsfacproc

Part of the Bioinformatics Commons, Genetics and Genomics Commons, and the Numerical Analysis and Scientific Computing Commons

Please take our feedback survey at: https://unomaha.az1.qualtrics.com/jfe/form/SV_8cchtFmpDyGfBLE

Recommended Citation

Cooper, Kathryn Dempsey; Currall, Benjamin; Hallworth, Richard; and Ali, Hesham, "An intelligent data-centric approach toward identification of conserved motifs in protein sequences" (2010). *Interdisciplinary Informatics Faculty Proceedings & Presentations*. 14.
https://digitalcommons.unomaha.edu/interdiscipinformaticsfacproc/14

This Conference Proceeding is brought to you for free and open access by the School of Interdisciplinary Informatics at DigitalCommons@UNO. It has been accepted for inclusion in Interdisciplinary Informatics Faculty Proceedings & Presentations by an authorized administrator of DigitalCommons@UNO. For more information, please contact unodigitalcommons@unomaha.edu.

An intelligent data-centric approach toward identification of conserved motifs in protein sequences

Kathryn Dempsey*[♦], Benjamin Currall[ס], Richard Hallworth[ס], Hesham Ali*[♦]
*College of Information Science and Technology, University of Nebraska at Omaha
[♦]Department of Pathology and Microbiology, University of Nebraska Medical Center
[ס]Department of Biomedical Sciences, Creighton University
Email: hali@mail.unomaha.edu

ABSTRACT

The continued integration of the computational and biological sciences has revolutionized genomic and proteomic studies. However, efficient collaboration between these fields requires the creation of shared standards. A common problem arises when biological input does not properly fit the expectations of the algorithm, which can result in misinterpretation of the output. This potential confounding of input/output is a drawback especially when regarding motif finding software. Here we propose a method for improving output by selecting input based upon evolutionary distance, domain architecture, and known function. This method improved detection of both known and unknown motifs in two separate case studies. By standardizing input considerations, both biologists and bioinformaticians can better interpret and design the evolving sophistication of bioinformatic software.

Keywords:

Protein sequences, intelligent tools, motif finding, prestin

1. INTRODUCTION

There are over 100 motif finding programs for DNA and protein sequences with no clear improvement from one to the next [3]. Despite multiple assessments showing that appropriate usage results in accurate motif detection, these programs' output can fail to identify known motifs or identify patterns with no functional implications in many laboratories [3,4]. One suggestion for this problem is that improper usage may be limiting program utility. Usage error can occur at any of the three stages: input preparation, algorithm execution, and output analysis. While algorithm execution and output analysis can often be improved by the designer of the program, input preparation is largely dependent upon the interface between the designers and individual users. It may be possible to better utilize motif finding software by standardizing the input expectations between the program designers and users. We investigated this proposal using two separate case studies:1) the well described voltage-gated potassium channels (Kv) family and 2) the poorly described solute carrier 26 (Slc26) family. Both were analyzed with motif finding software using a "traditional" approach and our proposed approach to determine if additional input consideration improved detection of motifs.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ACM-BCB 2010, Niagara Falls, NY, USA

Copyright © 2010 ACM ISBN 978-1-4503-0438-2... $10.00

1.1 Proposed approach

Our case studies focus on the input, rather than the algorithm, in motif finding software. To this end, we provide a comparison of a traditional approach versus our proposed approach. The traditional approach uses an "uninformed" dataset of homological sequences as input, whereas the proposed approach will use an "informed" dataset as shown in Figure 1.

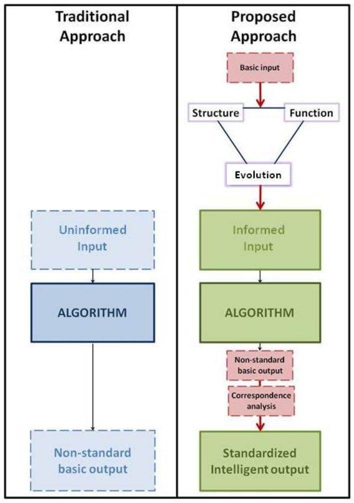


Figure 1: Traditional versus proposed model for dataset preparation in motif finding

1.2.1. Data preparation

The most straightforward method for data preparation is obtaining sequences homologous to the the protein(s) of interest. However, these datasets can often be biased by the availability of genomes (e.g. a preponderance of mammal and bacterial genomes). This often results in a dataset containing only very closely (i.e. mammals) and/or very distantly (i.e. bacterial) homologous sequences. Imbalances in consideration for phylogeny, function, and structure result in this uninformed, "traditional" input. Informed datasets will filter the uniformed dataset based upon evolution, structure, and function as described below.

1.2.1.1 Evolution

Though actual evolutionary rates (i.e. the molecular clock) can be estimated using measures such as molecular and paleontological dating, these methods are notorious for over- or under-estimating actual organism ages[1,2]. Until the true molecular clock can be quantified with more certainty, the user must rely on the intimate knowledge of their protein of interest and phylogenetic trees to determine evolutionary relationships. Once established, evolutionary relationships should be used to choose sequences so as not to bias towards a specific organism or clade while removing homologs that may have dissimilar structure-function from the protein of interest (see below).

1.2.1.2 Structure

Informed input also requires the identification of common structures (i.e. any known domains) within the dataset. Domains have their own structure-function relationship and independent analysis of each domain may reduce the possibility of false positives.

1.2.1.3 Function

Many homologs may have slightly or all together different function. Analyzing sequences with inherently different functions may obscure motifs for the protein of interest. Informed input should restrict sequences to those with predicted similar structure-function relationships.

1.2.2 Uniform motif scoring

Motif finding software output scoring is program dependent. In addition, motif finding programs allow for motifs of highly variable lengths, from as few as 4 to as many as 30 residues. This makes determining output quality unwieldy and thus a standardization technique for output and scoring serves as a means for method comparison.

1.3 Case studies

1.3.1 Voltage-gated potassium channels

One well described family of proteins is the voltage-gated potassium channels (Kv). The channel is composed of six transmembrane helices and a pore loop between helices five and six. These channels contain long cytoplasmic amino terminals. Subunits oligomerize into homo-/heterotetramers to form the functional channel [8]. There are two main domains conserved between voltage gated channel – the amino terminal T1(i.e. B2B) domain and the membrane bound ion transport domain as shown in Figure 3. The T1 domain participates in voltage-gated potassium channel tetramerization [8]. The residues that participate in tetramerization are distant in the primary structure and very somewhat within the voltage-gated potassium channel [8]. Thus, no well defined motif exists for tetramerization and it is not expected to be found by motif finding software.

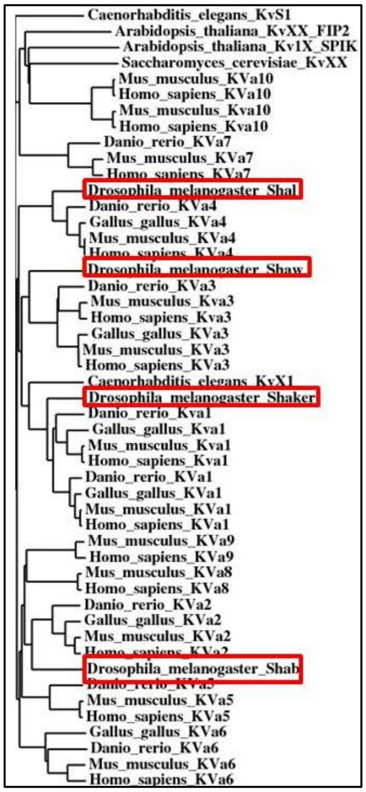


Figure 2: Phylogeny of all sequences in the uninformed voltage dataset with sequences in the informed datasets highlighted

The ion transport domain retains both the voltage sensor and the pore selectivity. The voltage sensor is mostly controlled by four conserved positive charges on the S4 transmembrane domain but distance between these residues suggest that the voltage sensor pattern is not necessarily expected to be detected by motif finding software. The pore selectivity, however, has a “signature sequence” (TxxTxGYG), which should be readily detectable by motif finding software [9].

1.3.2 Prestin and the SLC26 superfamily

The solute carrier 26 (Slc26) family of proteins is involved in diverse disease such as pendrin syndrome, cystic fibrosis, and adenoma. These proteins' function as anion transporters or channels [8,9], save for one exception: mammalian prestin. In mammals, the prestin (Slc26a5) ortholog acts as a motor protein, but in non-mammals, the prestin ortholog acts as an anion anti-porter[9,12]. This functional shift, at an evolutionary recent point, presents an interesting case study for many bioinformatics tools that examine structure-function relationship.

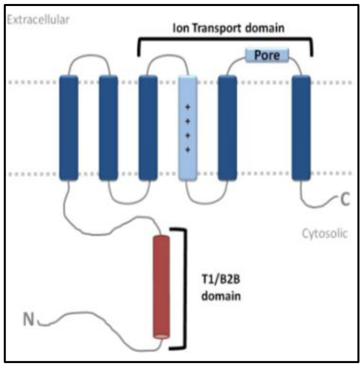


Figure 3: Structure of voltage-gated potassium channels.

Table 1: Description of voltage dataset

Dataset Name	Description	Size
Voltage Basic	Uniformed dataset	7 organisms 45 sequences 788.3 avg. length
Voltage T1	Informed Dataset 1 *D. melanogaster*; 4 proteins: Shab, Shaw, Shal, Shaker, conserved T1/B2B domain only	1 organism 4 sequences 176.3 avg. length
Voltage Ion Transport	Informed Dataset 2 *D. melanogaster*; 4 proteins: Shab, Shaw, Shal, Shaker, conserved ion transport domain	1 organism 4 sequences 94.3 avg. length

The Slc26 family structure is relatively unknown. Slc26 proteins are believed to have either 10 or 12 transmembrane domains as well as a relatively large carboxy terminal [6]. All Slc26s have a membrane bound xanthine uracil permease (XUP) and a carboxy terminal sulphate transporter anti-sigma factor antagonist (STAS) superfamily domain. There are no well defined motifs within these domains, however, a defined sulphate transporter motif is found to the amino side of the XUP domain [5]. It is unknown whether motif finding programs will detect any motifs within the Slc26 family.

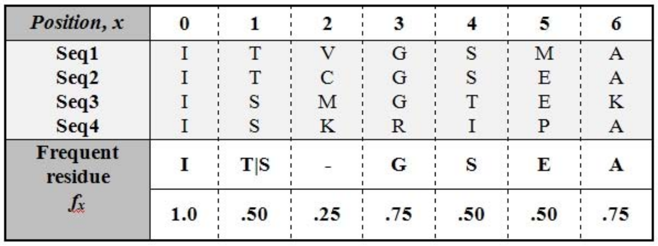
Table 2: Description of prestin dataset

Dataset Name	Description	Size
Prestin Basic	Uniformed Dataset	18 organisms 27 sequences 717.3 avg length
Prestin XUP	Informed Dataset 1 1 sequence per organism Non-mammalian XUP domain only	11 organisms 20 sequences 304.7 avg length
Prestin STAS	Informed Dataset 2 1 sequence per organism Non mammalian STAS domain only	11 organisms 20 sequences 166.1 avg length

2. METHODS

2.1 Data preparation

In both case studies motif finding was restricted to specific functional characteristics. In the Kv family, motifs for the A-type slow rectifiers were examined. In the Slc26 family, motifs for transport function were examined. For each case study, an uninformed and an informed dataset was input to the Gibbs Motif Sampler and the outputs were compared.



Position, x	0	1	2	3	4	5	6
Seq1	I	T	V	G	S	M	A
Seq2	I	T	C	G	S	E	A
Seq3	I	S	M	G	T	E	K
Seq4	I	S	K	R	I	P	A
Frequent residue	I	T|S	-	G	S	E	A
f_x	1.0	.50	.25	.75	.50	.50	.75

Figure 4: Example frequency table for found motifs

2.1.1 Voltage-gated potassium channels

Kv family protein sequences were obtained using NCBI's BLAST. Each sequence in the phylogenetic tree of Figure 3 was submitted as the uninformed dataset. An informed dataset was selected from a subset of these sequences (highlighted in red in Fig. 2) based on evolution, structure, and function. The shaker, shaw, shab, and shal proteins from *D. melanogaster* are approximately evolutionarily equidistant based on the phylogenetic tree and are thus included in the informed dataset. The dataset was further divided based upon the known domains of the Kv family. The informed dataset included shaker, shaw, shab, and shal proteins all have delayed rectifier and/or A-type function in contrast to the excluded KCNMA, which is a calcium gated Kv homolog and thus, was excluded from the informed dataset (Table 1).

2.1.2 Prestin SLC26 superfamily

Slc26 family protein sequences were obtained from NCBI's BLAST. These sequences were used for the uniformed dataset. For the informed dataset, 8 of the *H. sapiens* paralogs were used to remove bias towards any one particular paralog. The informed dataset was further divided based upon the XUP and STAS domains. Because mammalian prestin (Slc26a5) function as a motor protein, all orthologs were excluded from the informed dataset (Table 2).

2.2 Algorithm execution and parameters

It is important to acknowledge the actual execution of motif detection tools in our assessment. In Tompa *et al.* 2005, an assessment of approximately 10 popular motif detection programs was performed [6], and it was determined that, for general purposes, no extensive parameter tuning was necessary for optimal results. This was backed by the software authors, who contributed to the assessment by running their respective algorithms on the input as per request. We follow this sentiment by running the Gibbs Motif Sampler using default parameters in recursive mode over our datasets, which assumes 0+ sites per sequence and the original input for background training. Multiple runs were performed searching for motifs of varying lengths (6, 8, 10, 12, 14) in all datasets and each motif returned was scored sing an expectation value. Motif duplicates in length and content were removed from our final results.

2.3 Output analysis

We also propose a step in the traditional data pipeline called "Uniform Motif Scoring" (UMS) which uses an output preparation and algorithm to identify the strongest and shortest signals from found motifs. For a set of results from a motif detection program, it is a requirement for this approach that motifs can be represented as sets of characters in a gapless alignment with their sequence ID (Figure 4). Motif signals are identified by first examining the residue frequency at each position described as $f_x = CR_x/TR_x$, where:

CR_x is equal to the number of occurrences of <u>C</u>onsensus <u>R</u>esidue in position *x*,
TR_x is equal to the <u>T</u>otal number of <u>R</u>esidues in pos. *x*,
f_x is equal to the residue *f*requency at position *x*, or the ratio of MRR_x to TRR_x and
x is the position as defined by the initial alignment.

This allows us to create a consensus sequence (CS) with the frequency of the most represented residue at each position. We represent the consensus sequence CS by a set of frequencies where = {f_x, f_{x+1}, …, f_n}. The consensus frequency and sequence for our hypothetical example is highlighted in Figure 4. In addition, we show the CS at a variety of frequency thresholds and how it affects the content of the resulting CS.

2.3.1 Algorithm

We then find the motif (M) represented by the *longest continuous stretch* of frequencies *f* in C where all *f* are greater than or equal to *t*. The length of M also must be greater than or equal to 4 (though this value can be lowered if looking for shorter signals or single conserved residues). We present the following procedure to find M:

Input: *t*, C={f_x, f_{x+1},…, f_n}
Output: M = {m_x, m_{x+1},…, m_n}, the positions of the longest continuous stretch of positions in C where f_m are greater than or equal to *t*.

```
Let M={Ø}, TMP = {Ø}
for $i$ = 4 to $n$ do
    for $j$ = 0 to ($n - i + 1$) do
        for $k = j$ to ($j + i - 1$) do
            if $C_k < t$ then
                return $j = k + 1$;
                end;
            if $C_k \geq t$ then
                return TMP = TMP + $k$;
                $k$++;
            if size(TMP) ≥ size(M) ≥ $i$ then
                M = TMP;
    end
    return M;
end
```

Given C = {c_x, c_{x+1},…, c_n} and M = {m_x, m_{x+1},…, m_n}, we can define the character sequence of motif M and also the motif strength, M_s. To further enhance the sensitivity of the motif score, we also take into account the original amount of sequences input, $O_{s,}$ versus the amount of sequences returned that contained the motif result, F_s (Eq. 1):

$$M_s = \frac{\sum_1^n C_{(m_n)}}{n} * \frac{F_s}{O_s} \qquad (1)$$

3. RESULTS

3.1 Voltage gated potassium channel results

Gibbs Motif Sampler will return duplicates of the same motif with varying length (Figure 5). These returns were further analyzed using our Uniform Motif Scoring algorithm to determine the strongest signal within these duplicate motifs and were represented in table 3. The uniformly scored output from the uninformed Kv family dataset found only one significant motif (Table 3). This motif corresponds to the known Kv signature motif TxxTxGYG and was also identified as such by an ExPASY Prosite search. The signals found in the informed Kv family datasets had higher strength and corresponded to additional known motifs as found in ExPASYs Prosite database and Kv literature. The top results returned from the Gibbs runs on all three datasets before UMS are presented in sequence logo form in Figure 5 (logos made using [10]) and in text form in Table 3. If one motif was contained within another (i.e. motif was an extension of a shorter motif on either or both sides) the motif was still considered separate.

Table 3: Results for both Kv datasets after UMS.

		Scored Motif (t = 75%)	Motif Strength	Known Correspondence
Kv Uniformed	3	GYGD	0.935	Signature sub-seq.
Kv T1/B2B	1	EYFFDR	1.000	Located S4 region
	2	YYRTG	0.850	Located S4 region
	3	EYFFDR	1.000	Located S4 region
	4	NVGG	0.738	Located β1 region
	4	RHET	0.875	Located S4 region
Kv Ion Transport	1	TTVGYGD	0.964	Signature seq
	3	TMTTVGYG	0.969	Signature seq
	4	WFTFE	0.963	Located S2 region
	5	MTTVGYGDM	0.944	Signature seq

The Voltage B2B and Voltage Ion Transport datasets had slightly better performance. Two short motifs were found per dataset instead of one. In the Voltage Ion Transport dataset, the two motifs found were actually subsets of the Kv signature sequence motif (TxxTxGYG) and the S2 region motif which contains a negatively charged residue (E) critical for balancing positing charges in the membrane. Two different motifs were found in the B2B dataset, one with suggested structural importance.

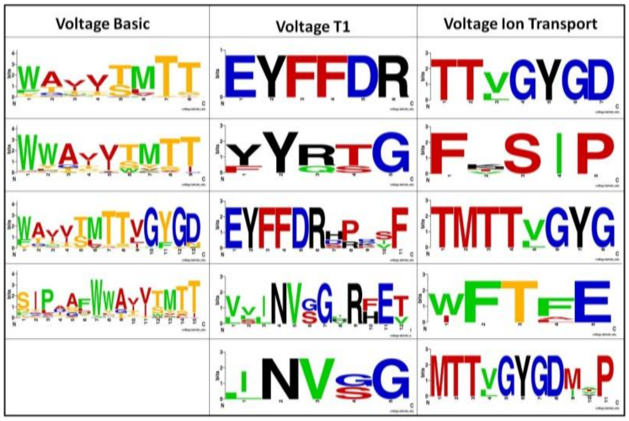


Figure 5: Gibbs results for both Kv datasets before UMS.

3.2 SLC26 superfamily and prestin

The Gibbs Motif Sampler found no significant motifs found in the uniformed dataset provided. The motifs found in the Prestin datasets for XUP and STAS are represented in Table 4. We identified 2 motifs in the XUP dataset and 2 motifs in the STAS dataset which correspond to known patterns in ExPASY's Prosite database shown in Table 4. In addition to the known correspondence of the unshortened motifs, motif 4 in the Prestin STAS dataset was identified as a potential Casein Kinase II phosphorylation site, raising more suspicions that this short conserved signal may be important for the universal structure and function in the SLC26 superfamily.

4. DISCUSSION

Here we have shown that the proposed model for preparation of data input can substantially improve the utility of motif finding software. The traditional approach yields very little (if any) substantial output, which could discourage the use of motif finding software altogether. Interestingly, these "informed" datasets are smaller than the traditional approach of inputting large sets of homologous sequences. This suggests that increasing the size of the dataset may actually reduce the viability of the output. Though algorithm issues are known to arise, the input rather than the algorithm dictated the viability of the output in our studies. The output viability required *a priori* knowledge of sequence evolution, structure, and function to determine the informed dataset. This *a priori* knowledge requires expertise from both the biological and informatic sciences which may further emphasize the need for common standards if continued successful integration of these disparate fields is to occur.

Table 4: Results for both prestin datasets after UMS.

		Scored Motif (t = 75%)	Motif Strength	Known Correspondence
Prestin Basic		No Motifs Found		
Prestin XUP	1	[M|S]L	-	-
	2	V[D|G][N|V]	-	-
	3	NQELI	-	N-myristoylation
	4	NQEL	-	N-myristoylation
	5	NQELIALG	--	N-myristoylation
Prestin STAS	1	DS[V|T]G	0.6667	Phosphorylation
	2	PIY[Y|F]AN	0.8000	C2K phosphoryl.
	3	[A|P]N[S|T]D[L|V]Y	-	-
	4	[S|T][I|V]HDA	0.6364	C2K phosphoryl
	5	D[S|T][V|S]G	0.7857	Phosphorylation

5. ACKNOWLEDGEMENTS

The authors would like to thank Dr. Dhundy Bastola for his help with this work. The project was funded in part by NIH grant P20 RR016469 from the INBRE program of the National Center of Research Resources.

6. REFERENCES

[1] Hedges SB, J Blair , M Venturi, J Shoe. *A molecular timescale of eukaryote evolution and the rise of complex multicellular life*. BMC Evolutionary Biology. 2004; **4**:2.

[2] Kumar S, Filipski A, Swarna V, Walker A, Hedges SB. *Placing confidence limits on the molecular age of the human-chimpanzee divergence*. Proceedings of the Natural Academy of Sciences. 27 Dec 2005; **102**(52):18842-18847.

[3] Quest D, K Dempsey, M Shafiullah, D Bastola, and H Ali. *MTAP: A Motif Tool Assessment Pipeline for Automated Assessment of De Novo Regulatory Motif Discovery Tool*. BMC Bioinformatics. 2008 Aug 12; **9** Suppl 9:S6.

[4] Tompa M, N Li, T Bailey , G Church , B DeMoor, E Eskin, A Favorov, M Frith, Y Fu, W Kent, V Makeev, A Mironov, W Noble, G Pavesi, G Pesole, M Regnier, N Simonis, S Sinha, G Thijs, J. van Helden, M Vandenbogaert, Z Weng, C Workman, C Ye, and Z Zhu. *Assessing Computational Tools for the Discovery of Transcription Factor Binding Sites*. Z Nature Biotechnology. 1 Jan 2005; **23**(1):137-144.

[5] Zheng, J., et al., *Prestin is the motor protein of cochlear outer hair cells.* Nature, 2000. **405**(6783): p. 149-55.

[6] Dorwart, M.R., et al., *The solute carrier 26 family of proteins in epithelial ion transport.* Physiology (Bethesda), 2008. **23**: p. 104-14.

[7] Yarov-Yarovoy V, Baker D, Catterall WA. *Voltage sensor conformations in the open and closed states in ROSETTA structural models of K(+) channels*. Proc Natl Acad Sci USA, 2006 May 9;103(19):7292-7. Epub 2006 Apr 28.

[8] Haitin Y, Yisharel I, Malka E, Shamgar L, Schottelndreier H, Peretz A, Paas Y, Attali B. *S1 constraints in the voltage sensor domain of Kv7.1 K+ channels*. PLoS One. 2008 Apr 9;3(4):e1935.

[9] Heginbotham K, Lu Z, Abramson T, MacKinnon R. *Mutations in the K+ channel signature sequence.* Biophys J. 1994 Apr; 66(4):1061-7.

[10]Crooks GE, Hon G, Chandonia JM, Brenner SE WebLogo: A sequence logo generator, Genome Res, 14:1188-1190, (2004)