ChemAlign: Biologically Relevant Multiple Sequence Alignment Using Physicochemical Properties

Hyrum Carroll[1], Mark Clement and Quinn Snell
CS Department, Brigham Young University
Provo, Utah, USA
HyrumCarroll@gmail.com,
{clement,snell}@cs.byu.edu

David McClellan
Bigelow Laboratory for Ocean Sciences
West Boothbay Harbor, Maine, USA
dmcclellan@bigelow.org

***Abstract*—We present a new algorithm, ChemAlign, that uses physicochemical properties and secondary structure elements to create biologically relevant multiple sequence alignments (MSAs). Additionally, we introduce the Physicochemical Property Difference (PPD) score for the evaluation of MSAs. This score is the normalized difference of physicochemical property values between a calculated and a reference alignment. It takes a step beyond sequence similarity and measures characteristics of the amino acids to provide a more biologically relevant metric. ChemAlign is able to produce more biologically correct alignments and can help to identify potential drug docking sites.**

***Keywords*-multiple sequence alignment; physicochemical properties; Physicochemical Properties Difference score;**

I. Introduction

Multiple sequence alignments (MSAs) are at the heart of several bioinformatics research areas. For example, alignments are used to identify conserved regions, which are crucial to finding drug docking sites. Current methods can miss biologically relevant features such as these because they only consider sequence similarity. Most of them are further limited because they do not incorporate secondary structure (SS) information. The globin family (with an average percent identity of 25.9% for the HOMSTRAD [1] data set) provides a good example of this in that it remains difficult for existing methods to align correctly. Previous algorithms align at best 38.4% of the positions correctly. Using physicochemical properties (PPs), ChemAlign correctly aligns 90.6% of the positions. Furthermore, as shown in Figure 1, regions determined from a ChemAlign alignment appear at a possible drug docking site.

ChemAlign uses PPs (e.g., volume, polarity and hydropathy) to produce biologically relevant alignments. Researchers have used these properties in various areas of bioinformatics [2]–[4]. Furthermore, they have varying effects depending on the SS where they occur. ChemAlign incorporates knowledge of the secondary structure elements (SSEs) (α-helices, β-strands and loops) to capitalize on this. Each amino acid in a protein belongs to one of the SSEs. Typically they are determined from tertiary structure information, if it is known, or are predicted. Protein SS has long been understood to be more conserved than the amino acid sequence [5]. Using this more resilient information has improved the accuracy of sequence alignments [6]–[8].

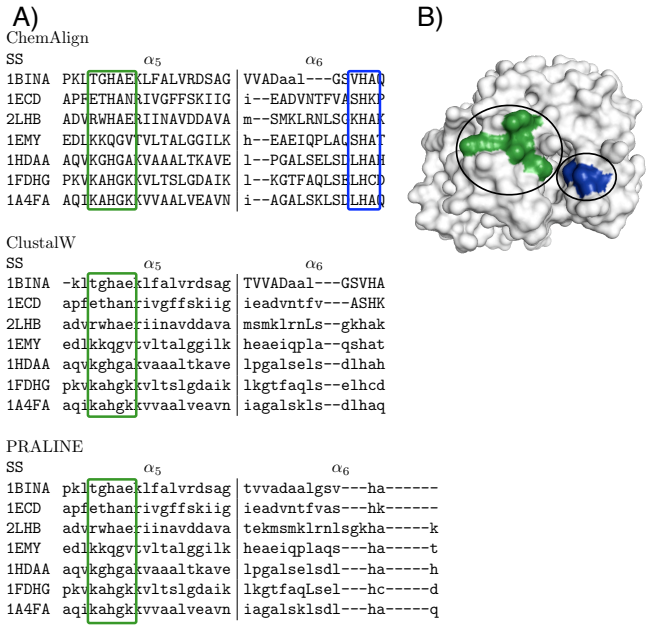

Figure 1. A) Alignments of the fifth and sixth α-helices of the globin data set. Uppercase amino acids match the reference data set. ChemAlign is able to align the vast majority (90.6%) of the positions correctly, compared to ClustalW (38.4%) and PRALINE (24.4%). ChemAlign is able to find both of them, while ClustalW and PRALINE only find the first one. B) Example globin protein, Hemoglobin (1A4FA), with highlighted conserved regions corresponds with conserved regions. These regions are at a possible drug docking site.

In this paper, we explore the hypothesis that using physicochemical properties and secondary structures produces biologically relevant multiple sequence alignments. To do so, we introduce ChemAlign, which incorporates both physicochemical properties and secondary structures.

II. Related Work

MSA algorithms that are related to ChemAlign fit into three categories. First, algorithms like MAFFT [9] and ProbCons [10] that also use primary sequence information. In a recent benchmarking study [11], these two applications performed the best. Second, algorithms that incorporate SSEs, like PRALINE [6]. PRALINE is a MSA algorithm that builds an alignment without SS information, then iterates between predicting the SSEs and building an

[1]Current Address: National Institutes of Health, Bethesda, MD, USA

alignment. It is also subject to the same limitation as the primary sequence alignment algorithms—that of not being able to correctly produce alignments governed by PPs. Additionally, PRALINE is only available through an interactive website and therefore requires substantial amounts of human interaction for large-scale use or testing. Third, algorithms that integrate PPs. While researchers are using PPs for various analyses, few have incorporated them into sequence alignment. Those that do, use them in pairwise alignments [3], to find matching subsequences [4], and to adjust gap penalties [12]. ChemAlign extends these ideas to produce MSAs.

III. Methods

ChemAlign is a multiple sequence alignment algorithm that uses the physicochemical property values and secondary structures of amino acids. It employs a traditional dynamic programming approach during both the pairwise and the progressive phases. After calculating all of the pairwise "distances" between sequences, ChemAlign clusters them to produce a neighbor-joining guide tree [13]. This tree directs the order that sequences and alignments of sequences are aligned in the progressive stage. ChemAlign also uses affine gap penalties. Instead of using a substitution matrix based solely on log-odds probabilities from an amino acid database, ChemAlign combines amino acid exchange counts with normalized differences of PPs. Additionally, different substitution matrices are employed for different SSEs. In the rest of this section, we explain ChemAlign's use of PPs and SSs and how it calculates gap costs.

A. Substitution Matrices

ChemAlign uses a substitution matrix comprised of both observed amino acid exchanges and differences between PPs. To obtain the observed amino acid exchanges, we built a reference database of alignments with their SSs. We combined the OXBench database [14] with the respective SSs from the RCSB Protein Data Bank (PDB) [15]. Only those sequences in OXBench that had an exact match with amino acid sequences in the PDB were included. We counted the number of each set of amino acid pairs for each possible SSE pair producing four matrices of observed amino acid exchanges: $O^{\alpha}, O^{\beta}, O^{l}$, and O^{m} (m stands for mismatch). These matrices are combined with the normalized difference matrix D^{p} (for a PP p) using Equation 1.

$$D^{p}_{i,j} = 1 - \frac{2 * |PP[i] - PP[j]|}{argmax_x(PP[x]) - argmin_y(PP[y])} \quad (1)$$

Here, i and j are amino acids. The values of D^{p} range from -1.0 for the most dissimilar pair of amino acids to 1.0 for identical amino acids. For this work, we use the Effective Partition Energy [16] for its aggregate characteristics as an illustrative PP. This property includes hydrophobic, hydrogen bonding and electrostatic energies. Each of the O matrices are multiplied element-wise with D^{p} to get $M^{\alpha}, M^{\beta}, M^{l}$, and M^{m}. Combining the O matrices with D^{p} aggregates the benefits of each. Finally, the log-odds probabilities of the values in each of the M matrices are calculated to get the substitution matrices S^{α}, S^{β}, S^{l} and S^{m}:

$$S_{i,j} = log\left(\frac{l_{i,j}}{f_i f_j}\right) \quad (2)$$

Here, $l_{i,j}$ is the likelihood that amino acids i and j appear aligned in the database and f_i is the background frequency of amino acid i.

	α-helix	β-strand	loop
α-helix	7.11		
β-strand	-12.81	2.97	
loop	-2.42	-3.33	1.95

Figure 2. Secondary structure scoring matrix N. The values are log-odd ratios based on observed counts in the OXBench-PDB database.

B. Incorporating Secondary Structure

ChemAlign uses a straightforward approach to incorporate protein SSs into both pairwise and progressive alignment. The SS influences the alignment in two ways: first, the choice of a substitution matrix and second, an additional score for (mis)matching of the SSEs. First, ChemAlign uses a substitution matrix according to the SSEs of the two amino acids currently being considered. If the SSEs are the same, then the S^{α}, S^{β} or S^{l} matrix is used, otherwise, the mismatch matrix, S^{m}, is used. ChemAlign also incorporates SSs by adding a (mis)match score for the SSEs to the (mis)match score for the amino acids. The SSE scores are specified by the matrix $N_{c,d}$ (where c and d are SSEs) as shown in Figure 2. N is the log-odds ratios of the observed matches of the SSEs in the OXBench-PDB database. Incorporating N aligns SSs, which are typically more conserved than the amino acids themselves [5].

C. Reference Sum of Pairs Score

A commonly applied metric for MSA algorithms is the reference sum of pairs score–the percentage of positions in a calculated alignment that match the same character in a reference alignment:

$$\frac{1}{nq}\sum_{i}^{n}\sum_{k}^{q}\delta(s_i(k), r_i(k)) \quad (3)$$

Where $s_1, \ldots s_n$ are sequences of length l from a calculated alignment, $r_1, \ldots r_n$ sequences of length p from a reference alignment, $q = min(l, p)$ and δ is an identity function.

D. Physicochemical Property Difference (PPD) Score

In addition to using the reference sum of pairs score, we also look at the normalized difference in PPs values, or the PPD score. The score is calculated as follows:

$$\frac{1}{nq}\sum_{i}^{n}\sum_{k}^{q} D^{p}_{s_i(k), r_i(k)} \quad (4)$$

PPD scores range from a theoretical minimum of -1.0 to 1.0. In general, a negative PPD score means that the

average amino acid pairing in an alignment is worse than the average difference in the PP values. A score of 1.0 means the calculated alignment is the same as the reference alignment. This score takes a step beyond sequence similarity and measures characteristics of the amino acids.

E. Experimental Setup

To analyze the accuracy of ChemAlign, we looked at three databases of reference MSAs: BAliBASE [17], HOMSTRAD [1], and SMART [18]. We combine each of the sequences in the databases with the SSEs from the PDB. Only Sequences with a perfect sequence match in the PDB were included.

For ChemAlign, we used the following command: `ssalign( subMatA=`S^{α} `subMatB=`S^{β} `subMatL=`S^{l} `subMat=`S^{m} `ss=<SSE file>)`. These arguments specify files containing the substitution matrices S^{α}, S^{β}, S^{l} and S^{m} and a file containing the SSEs defined by DSSP [19]. We used the default arguments for the following programs: ClustalW (version 2.06); MAFFT (6.240); and ProbCons (1.12).

IV. Results

To quantitatively assess the performance of ChemAlign, its accuracy was compared with that of ClustalW, MAFFT and ProbCons. These programs were chosen for their performance, ubiquity and ease of use [11]. Both the reference sum of pairs score and the PPD scores were used in our evaluation. An analysis of three databases and an in-depth look at the globin domain family are presented. In summary, ChemAlign achieves comparable or higher accuracy scores and a more biologically meaningful alignment than the other programs tested.

A. Comparison of Algorithms on Difficult Data Sets

In order to examine the impact of PP alignments, sixteen data sets were selected from the BAliBASE, HOMSTRAD, and SMART databases that have the lowest sequence identity and a large number of sequences. The alignments produced by ClustalW, MAFFT and ProbCons were compared to ChemAlign using the reference sum of pairs metric (see Figure 3). A value of zero on the vertical axis indicates that the performance was identical. A score of 1.0 indicates that ChemAlign got all of the positions right and the other algorithm got all positions wrong. Negative vertical axis values indicate that the competing algorithm achieved better alignments than ChemAlign. Many of the ChemAlign alignments are significantly better than the competing algorithms. When the other algorithms are superior, the alignment produced by ChemAlign is generally close in score.

B. Reference Sum of Pairs Scores for All Data Sets

Table I reports the mean reference sum of pairs score for ChemAlign, ClustalW, MAFFT, and ProbCons on the BAliBASE, HOMSTRAD, and SMART databases. ChemAlign achieves comparable mean reference sum of pairs scores to the other methods tested. It consistently obtains

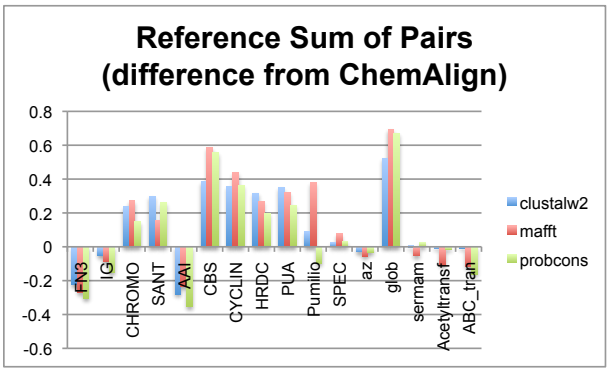

Figure 3. Differences between competing algorithms and ChemAlign. ChemAlign finds significantly better alignments for difficult data sets.

Table I
Mean Reference Sum of Pairs Scores

Database	ChemAlign	ClustalW	MAFFT	ProbCons
BAliBASE	37.8%	38.7%	37.1%	39.7%
HOMSTRAD	58.8%	59.5%	56.3%	59.7%
SMART	59.8%	58.6%	56.9%	59.9%

scores higher than MAFFT, scores higher than ClustalW on the SMART database and higher than ProbCons on the HOMSTRAD database. Many of these data sets have few sequences, or have high sequence similarity, so the differences between algorithms are less pronounced.

C. Physicochemical Property Difference Scores

We also evaluated the alignments generated from ChemAlign, ClustalW, MAFFT and ProbCons using the PPD score (with the PP Effective Partition Energy). ChemAlign achieves similar or superior PPDs scores, suggesting that the alignments are equally or more biologically accurate. While the Effective Partition Energy generally captures the forces of mutation here, researchers can also use the PPD score to evaluate additional properties (i.e., polarity or volume) affecting their alignments.

D. Globin Domain Alignment

The globin data set, used here as an example, was taken from the HOMSTRAD database, and is composed of 41 protein sequences, all of which have representative crystal structures in the PDB. Seven different categories of globin proteins are represented in this data set. Such protein diversity, in terms of primary and SS, as well as overall function, makes accurate alignment notoriously difficult.

The globin data set has a low percent identity of 25.9%, making it difficult for current methods to correctly align. ChemAlign is able to get 90.6% of the positions correct, while MAFFT only achieves 21.2% of them correct (ClustalW: 38.4%, ProbCons: 23.6% and PRALINE: 24.4%). In terms of percentages, ChemAlign is between 135.9–328.8% better (3,727–4,951 more positions) than the other methods. ChemAlign earns a PPD score of 0.79, which is between 76.2–242.6% better than the other methods. These scores reflect that ChemAlign produces alignments with columns of higher Effective Partition Energy similarity than the other algorithms. This is a characteristic of biologically relevant alignments.

Table II
PHYSICOCHEMICAL PROPERTY DIFFERENCE SCORES

Database	ChemAlign	ClustalW	MAFFT	ProbCons
BAliBASE	0.400	0.400	0.369	0.376
HOMSTRAD	0.632	0.635	0.606	0.628
SMART	0.643	0.632	0.610	0.629

ChemAlign is able to correctly align the vast majority of the amino acids throughout the globin data set. ClustalW only aligns the first, part of the second and the third α-helices correctly. PRALINE correctly aligns only the first of the eight α-helices. Figure 1 shows the ChemAlign, ClustalW and PRALINE alignments of the fifth and sixth α-helices. Highlighted on the alignments are the most conserved regions (using a sliding window of size three). ChemAlign is able to find both regions, while ClustalW and PRALINE only find the first one. The positions of these regions on the protein is a potential drug docking site. Alignment methods that do not incorporate PPs and SS information can limit the discovery of such regions.

V. CONCLUSION

Multiple sequence alignments are the foundation for several bioinformatics research areas. For example, identifying genes for drug development relies on an accurate alignment of sequences. Current methods struggle to accurately align data sets with low percent identity. ChemAlign is a new algorithm that addresses this problem by using a physicochemical property to produce biologically relevant MSAs. It also incorporates SSEs to overcome limitations employed by traditional approaches that use the "'average' site in the 'average' protein" [2]. Leveraging this additional information, it is able to find more potential drug docking sites than other algorithms (see Figure 1). Additionally, we introduce the Physicochemical Property Difference (PPD) score. This score measures the average difference in values for a physicochemical property for all pairs of amino acids in an alignment. It takes a step beyond sequence similarity and measures characteristics of the amino acids. ChemAlign achieves comparable or superior PPD scores than the other algorithms tested.

ChemAlign is implemented in the software package PSODA [20]. PSODA is free and available for several operating systems at http://dna.cs.byu.edu/psoda.

VI. FUTURE WORK

We are improving ChemAlign by extending the difference in PPs matrix, D, to handle multiple properties with weights. Additionally, we are looking at increasing the specificity of the substitution matrices by using different PPs for each of the SSs.

REFERENCES

[1] K. Mizuguchi, C. M. Deane, T. L. Blundell, and J. P. Overington, "HOMSTRAD: A database of protein structure alignments for homologous families," *Protein Science*, vol. 7, no. 11, pp. 2469–2471, 1998.

[2] J. L. Thorne, N. Goldman, and D. T. Jones, "Combining protein evolution and secondary structure," *Molecular Biology and Evolution*, vol. 13, no. 5, pp. 666–673, 1996.

[3] P. Gonnet and F. Lisacek, "Probabilistic alignment of motifs with sequences," *Bioinformatics*, vol. 18, no. 8, pp. 1091–1101, 2002.

[4] K. Gupta, D. Thomas, S. Vidya, K. Venkatesh, and S. Ramakumar, "Detailed protein sequence alignment based on Spectral Similarity Score (SSS)," *BMC Bioinformatics*, vol. 6, no. 105, 2005.

[5] J. Gibrat, T. Madej, and S. Bryant, "Surprising similarities in structure comparison," *Current Opinion in Structural Biology*, vol. 6, no. 3, pp. 377–385, 1996.

[6] J. Heringa, "Two strategies for sequence comparison: profile-preprocessed and secondary structure induced multiple alignment," *Comput. Chem*, vol. 23, pp. 341–364, 1999.

[7] A. Jennings, C. Edge, and M. Sternberg, "An approach to improving multiple alignments of protein sequences using predicted secondary structure," *Protein Engineering Design and Selection*, vol. 14, no. 4, pp. 227–231, 2001.

[8] X. Zhang and T. Kahveci, "A new approach for alignment of multiple proteins," *Proceedings of the 11th Pacific Symposium on Biocomputing*, pp. 339–350, 2006.

[9] K. Katoh, K. Kuma, H. Toh, and T. Miyata, "MAFFT version 5: improvement in accuracy of multiple sequence alignment," *Nucleic Acids Research*, vol. 33, no. 2, pp. 511–518, 2005.

[10] C. B. Do, M. S. P. Mahabhashyam, M. Brudno, and S. Batzoglou, "ProbCons: probabilistic consistency-based multiple sequence alignment," *Genome Research*, vol. 15, pp. 330–340, 2005.

[11] H. Carroll, W. Beckstead, T. O'Connor, M. Ebbert, M. Clement, Q. Snell, and D. McClellan, "DNA Reference Alignment Benchmarks Based on Tertiary Structure of Encoded Proteins," *Bioinformatics*, vol. 23, no. 19, pp. 2648–2649, 2007.

[12] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice," *Nucleic Acids Research*, vol. 22, pp. 4673–4680, 1994.

[13] N. Saitou and M. Nei, "The neighbor-joining method: a new method for reconstructing phylogenetic trees," *Molecular Biology and Evolution*, vol. 4, pp. 406–425, 1987.

[14] G. P. S. Raghava, S. M. J. Searle, P. C. Audley, J. D. Barber, and G. J. Barton, "OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy," *BMC Bioinformatics*, vol. 4, no. 47, 2003.

[15] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2000.

[16] S. Miyazawa and R. L. Jernigan, "Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation," *Macromolecules*, vol. 18, no. 3, pp. 534–552, 1985.

[17] J. D. Thompson, P. Koehl, R. Ripp, and O. Poch, "BAliBASE 3.0 latest developments of the multiple sequence alignmnet benchmark," *Proteins: Structure, Function, and Bioinformatics*, vol. 61, no. 1, pp. 127–136, 2005.

[18] I. Letunic, R. R. Copley, S. Schmidt, F. D. Ciccarelli, T. Doerks, J. Schultz, C. P. Ponting, and P. Bork, "SMART 4.0: towards genomic data integration," *Nucleic Acids Research*, vol. 32, pp. D142–D144, 2004.

[19] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, 1983.

[20] H. Carroll, A. R. Teichert, J. Krein, K. Sundberg, Q. Snell, and M. Clement, "An open source phylogenetic search and alignment package," *International Journal of Bioinformatics Research and Applications*, vol. 5, pp. 349–364, 2009.