Sequence context-specific profiles for homology searching

A. Biegert and J. Söding[1]

Gene Center Munich and Center for Integrated Protein Science, Ludwig Maximilian University of Munich, Feodor-Lynen-Strasse 25, 81377 Munich, Germany

Edited by David Baker, University of Washington, Seattle, WA, and approved January 9, 2009 (received for review October 24, 2008)

Sequence alignment and database searching are essential tools in biology because a protein's function can often be inferred from homologous proteins. Standard sequence comparison methods use substitution matrices to find the alignment with the best sum of similarity scores between aligned residues. These similarity scores do not take the local sequence context into account. Here, we present an approach that derives context-specific amino acid similarities from short windows centered on each query sequence residue. Our results demonstrate that the sequence context contains much more information about the expected mutations than just the residue itself. By employing our context-specific similarities (CS-BLAST) in combination with NCBI BLAST, we increase the sensitivity more than 2-fold on a difficult benchmark set, without loss of speed. Alignment quality is likewise improved significantly. Furthermore, we demonstrate considerable improvements when applying this paradigm to sequence profiles: Two iterations of CSI-BLAST, our context-specific version of PSI-BLAST, are more sensitive than 5 iterations of PSI-BLAST. The paradigm for biological sequence comparison presented here is very general. It can replace substitution matrices in sequence- and profile-based alignment and search methods for both protein and nucleotide sequences.

alignment | pseudocounts | substitution matrix | context-sensitive

Substitution matrices quantify the similarity between amino acids or nucleotides (1–3). As a mainstay of biological sequence comparison, they are at the heart of standard alignment methods such as the Needleman–Wunsch and Smith–Waterman algorithms (4, 5), which find the alignment with the maximum sum of similarity scores between aligned residues or bases. Sequence-search programs such as BLAST and FASTA (6, 7) use substitution matrices to score short seeds and final alignments, multiple alignment programs such as CLUSTALW (8) employ them in sum-of-pairs scoring to quantify the similarity between aligned sequence-profile columns, and in sequence profile-based methods such as PSI-BLAST (9) or HHsearch (10) they are used for calculating pseudocounts (11, 12).

For proteins, the importance of substitution matrices to identify homologs and calculate accurate alignments has stimulated various advances. Yu *et al.* (13) have developed a rationale for compositional adjustment of amino acid substitution matrices by transforming the background frequencies implicit in a substitution matrix to frequencies appropriate for the comparison of protein sequences with nonstandard global amino acid composition. Others have derived specialized transmembrane substitution matrices from alignments of experimentally verified or predicted transmembrane segments to improve alignments of sequences with transmembrane regions (14–16). The logic is that the structural environment of an amino acid residue partly influences into what amino acids it is likely to mutate.

Taking this idea a step further, so-called structure-dependent substitution matrices have been trained for a number of environments, defined by a combination of secondary structure state, solvent accessibility class, environmental polarity class, and/or hydrogen bonding (17–20). EvDTree (21) also computes structure-dependent substitution scores, but the selected structural descriptors depend on residue types. All of these structural environment-dependent matrices allow for the detection of more homologous proteins than standard substitution matrices. However, their application is limited by the need to know the structure of one of the proteins to be compared.

In contrast, sequence context-dependent methods do not rely on 3D structure information to define local environments. They describe the environment of a residue by the sequence surrounding it. Jung and Lee trained several 400 × 400 substitution matrices for contexts consisting of pairs of residues up to 4 positions apart and obtained a 30% increase in sensitivity on a set of 107 proteins (22), although this result could not be confirmed in a large-scale study (23). Gambin *et al.* derived 400 substitution matrices, one for each context consisting of the 2 residues neighboring the central residue (24, 25). PHYBAL (26) models the selective pressure inside and outside of hydrophobic blocks by 2 different substitution matrices and 2 different sets of gap penalties.

Huang *et al.* (27) took a decisive step forward, employing 281 substitution matrices for 281 states of a hidden Markov model (HMM) trained on sequences of known structure. Each HMM state represents a single profile column. Context information is encoded essentially in the transition probabilities between the states. By mixing mutation probabilities from the substitution matrices, weighted by posterior probabilities for the corresponding HMM states, HMMSUM achieved considerable improvements in alignment quality when compared with standard substitution matrices (27). We expect such sequence contexts to predict mutation probabilities better than structural environments, because very different sequences with very specific amino acid preferences can adopt similar local structures (28). When all of these sequences are pooled into the same structural environment, the specific amino acid preferences are lost.

In this work, we present a new method that derives sequence context-specific amino acid similarities from 13-residue windows centered on each residue. We predict the expected mutation probabilities for each position by comparing its sequence window to a library with thousands of *context profiles*, generated by clustering a large, representative set of sequence-profile windows. The mutation probabilities are obtained by weighted mixing of the central columns of the most similar context profiles (see Fig. 1*B*). Whereas iterative profile search tools such as PSI-BLAST align homologous, long sequence matches to the query with weights independent of the quality of the match, our method aligns mostly nonhomologous, ungapped, short profiles, giving higher weights to better matching profiles. In contrast to HMMSUM, no substitution matrices are needed. Also, the context information is encoded explicitly in the context profiles

Author contributions: J.S. designed research; A.B. performed research; A.B. contributed new reagents/analytic tools; A.B. analyzed data; and A.B. and J.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. E-mail: soeding@lmb.uni-muenchen.de.

This article contains supporting information online at www.pnas.org/cgi/content/full/0810767106/DCSupplemental.

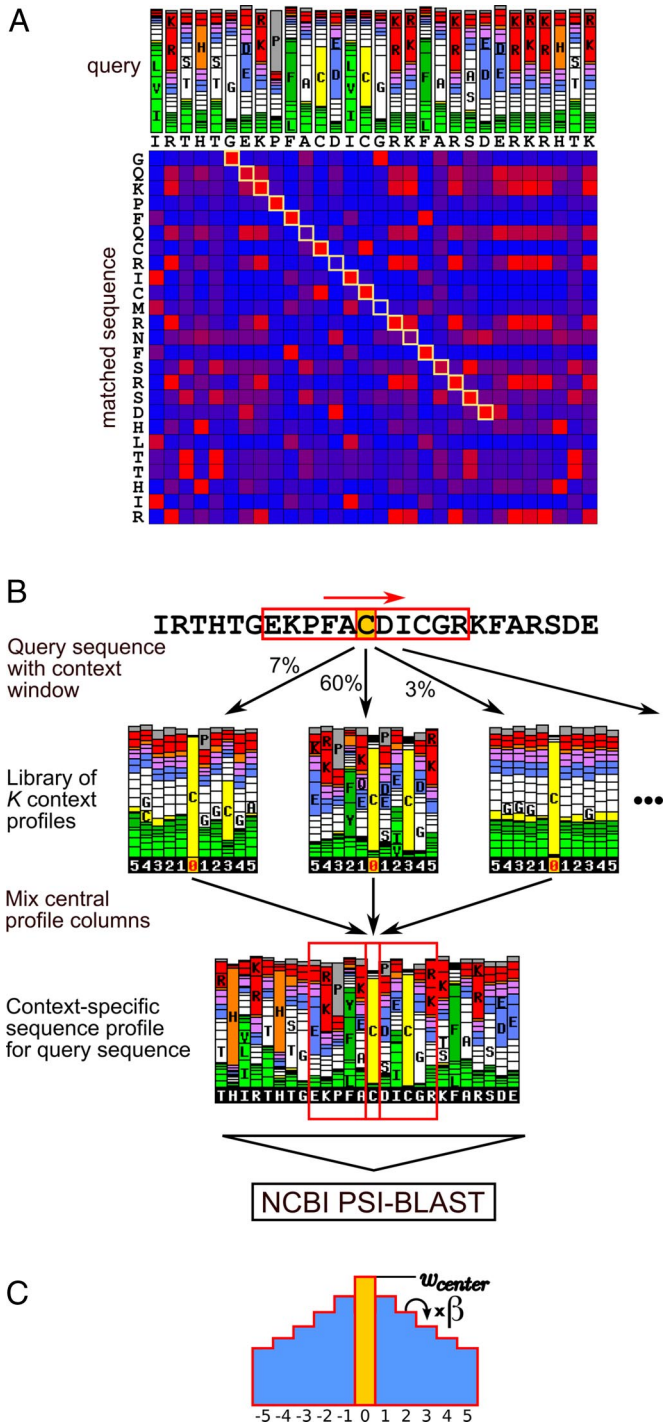

Fig. 1. Method of context-specific sequence comparison. (*A*) Sequence search/alignment algorithms find the path that maximizes the sum of similarity scores (color-coded blue to red). Substitution matrix scores are equivalent to profile scores if the sequence profile (colored histogram) is generated from the query sequence by adding artificial mutations with the substitution matrix pseudocount scheme. Histogram bar heights represent the fraction of amino acids in profile columns. (*B*) Computation of context-specific pseudocounts. The expected mutations (i.e., pseudocounts) for a residue (highlighted in yellow) are calculated based on the sequence context around it (red box). Library profiles contribute to the context-specific sequence profile with weights determined by their similarity to the sequence context (see percentages). The resulting profile can be used to jump-start PSI-BLAST, which will then perform a sequence-to-sequence search with context-specific amino acid similarities. (*C*) Positional window weights are chosen to decrease exponentially with the distance from the center position to model the decreasing information value of farther positions for the central profile column.

with no need for transition probabilities. This leads to a simpler computation and to a much better runtime that scales linearly instead of quadratically with the number of states/contexts (see *Discussion*). The context library can therefore be many times larger, and hence finer-grained, than in HMMSUM, enabling us to describe contexts as specific as "a large aliphatic residue with preference for I or M on the hydrophobic face of an amphipathic α-helix," for example.

A crucial insight for achieving speeds comparable to substitution matrix-based methods such as BLAST is this: Sequence-to-sequence comparison by using a substitution matrix is exactly equivalent to profile-to-sequence comparison, if the sequence profile is calculated from one of the sequences by using full substitution matrix pseudocounts. Hence, we can employ profile-based methods, which have similar speeds as their sequence-based counterparts, to implement sequence context-specific amino acid similarities.

CS-BLAST, our context-specific version of BLAST, works in the following way. We generate a sequence profile for the query sequence by using context-specific pseudocounts and then jump-start NCBI's profile-to-sequence search method PSI-BLAST with this profile. We demonstrate that, on a difficult benchmark set, sequence searches with our new context-specific amino acid similarities are more than twice as sensitive as BLAST with the standard BLOSUM62 substitution matrix, produce higher-quality alignments, and generate reliable E-values, all without loss of speed.

Finally, we apply the new paradigm to profile-to-sequence comparison by calculating context-specific pseudocounts for sequence profiles. The only difference to the previously described sequence-based scheme is that we now compare sequence-*profile* windows with our library of context profiles. In contrast to substitution matrix and Dirichlet pseudocounts (11, 12, 29, 30), these pseudocounts do not depend only on the single-profile column, but also on the entire sequence context of the profile column. We report considerable improvements of this context-specific scheme (CSI-BLAST) over PSI-BLAST.

Results

We first show that amino acid substitution scores are directly related to pairwise amino acid mutation probabilities and sequence-profile pseudocounts. We can therefore derive sequence context-specific amino acid similarity scores from context-specific mutation probabilities. These mutation probabilities can be predicted with a probabilistic model by using a large library of sequence-profile windows representing very specific local sequence contexts.

Any matrix of substitution scores $S(x, y)$ describing the similarity between amino acids x and y can be written in the form (31) $S(x, y) = \text{const} \times \log [P(x, y)/P(x)P(y)]$, where $P(x, y)$ is the probability that x and y occur aligned to each other in an alignment of homologous sequences, and $P(x)$ and $P(y)$ are the *background probabilities* of x and y to occur in representative sequences (whether aligned or unaligned). This can also be written as a log odds score, $S(x, y) = \log [P(y|x)/P(y)]$, where $P(y|x) = P(x, y)/P(x)$ is the conditional probability of y given x, i.e., the probability for amino acid x to mutate into y. If y occurs more often in positions aligned with an x (described by $P(y|x)$) than what would be expected by chance (described by $P(y)$), then the score is positive, otherwise negative.

We next explore the connection of mutation probabilities $P(y|x)$ with sequence-profile pseudocounts. A *sequence profile* is a matrix $p(i, y)$ that succinctly represents a multiple alignment of homologous sequences: $p(i, y)$ is the frequency of amino acid y in column i of the multiple alignment. The profile describes what amino acids are likely to occur in related sequences at each position, or, in other words, the probability of a residue at position i to mutate into amino acid y. A single sequence (x_i) can

be turned into a sequence profile by adding artificial mutations (i.e., *pseudocounts*) with the method of substitution matrix pseudocounts (11, 12): $p(i, y) = P(y|x_i)$. Here, $P(y|x_i)$ are the conditional probabilities giving rise to substitution matrix $S(x, y)$. The profile-to-sequence score of column i of this single-sequence profile p with residue y_j of a sequence (y_j) is

$$S(p(i,\cdot), y_j) := \log \frac{p(i, y_j)}{P(y_j)} = \log \frac{P(y_j|x_i)}{P(y_j)} = S(x_i, y_j). \quad [1]$$

Hence, substitution matrix scores can be seen as a special case of profile-to-sequence scores, where the profile is generated from one of the sequences by using substitution matrix pseudocounts.

Fig. 1*A* illustrates the equivalence of sequence-to-sequence and profile-to-sequences scoring with the alignment matrix of 2 zinc-finger sequences (x_i) and (y_j). The query profile resulting from the artificial mutations is illustrated as a histogram, in which the bar heights are proportional to the corresponding amino acid probabilities $p(i, y)$. The score of each matrix cell (i, j) can be interpreted in 2 ways: either as sequence-to-sequence score $S(x_i, y_j)$ between residues x_i and y_j, or as profile-to-sequence score $S(p(i,\cdot),y_j)$ between profile column $p(i,\cdot)$ and residue y_j.

In the above schemes, the expected mutation probabilities $P(y/x_i)$ at position i depend only on the single amino acid x_i. However, the sequence context X_i, defined below, contains much more information than just residue x_i itself about what amino acids to expect in related sequences. If we were able to calculate a context-specific mutation probability $P(y|X_i)$, we could define a score in a way analogous to Eq. **1**, but by using a *context-specific* profile $p_{cs}(i, y) = P(y|X_i)$ instead of $P(y|x_i)$.

The context X_i is defined as the window of l residues surrounding x_i, i.e., $X_i = (x_{i-d}, \ldots, x_{i+d})$ with $l = 2d + 1$. To predict the mutation probabilities for each position i, we compare its sequence window X_i with a precomputed library of K context profiles, $p_1, \ldots, p_K$, each of length l. The context-specific mutation probability $P(y|X_i)$, i.e., the probability of observing amino acid y in a homologous sequence given context X_i, will be calculated by a weighted mixing of the amino acids in the central columns of the most similar context profiles (Fig. 1*B*). To derive the weight of each profile p_k, we first need the probability $P(X_i|p_k)$ that the sequence window X_i is *emitted* by profile p_k, which is equal to the product of probabilities for x_{i+j} $(j \in \{-d, \ldots, d\})$ being emitted by profile column $p_k(j,\cdot)$: $P(X_i|p_k) = \Pi_{j=-d}^{d} p_k(j, x_{i+j})$. Because the inner positions in the window will be most informative to predict the amino acid distribution for the central residue, we can refine the above formula by defining coefficients w_j, which weight the contribution of each window position: $P(X_i|p_k) \propto \Pi_{j=-d}^{d} p_k(j, x_{i+j})^{w_j}$. The values of w_j are parameterized by w_{center} and β (see Fig. 1*C*). (For i within d residues from either end of (x_i) the product runs only over those j for which x_{i+j} is defined.)

Next, we need to know the probability $P(p_k|X_i)$ that profile p_k was the one that emitted X_i. Using Bayes' theorem, we find

$$P(p_k|X_i) = \frac{P(X_i|p_k)\, P(p_k)}{P(X_i)} \propto P(p_k) \prod_{j=-d}^{d} p_k(j, x_{i+j})^{w_j}. \quad [2]$$

$P(p_k)$ is the Bayesian *prior probability* for profile p_k, determined in the process of computing the profile library [supporting information (SI) *Appendix*]. It quantifies the probability that a sequence window is emitted by profile p_k *prior* to knowing that sequence window. $P(X_i) = \Sigma_k P(X_i|p_k)\, P(p_k)$ is a normalization constant.

We can now calculate the context-specific mutation probabilities $P(y|X_i)$ by mixing the amino acid distributions $p_k(0, y)$ from the central columns of all K profiles with weights $P(p_k|X_i)$:

$$P(y|X_i) \propto \sum_{k=1}^{K} p_k(0, y) P(p_k|X_i). \quad [3]$$

Normalizing over all 20 amino acids yields the expected mutation probability $P(y|X_i)$. To have more flexibility in adjusting the diversity of the context-specific profile $p_{cs}(i,\cdot)$, we mutate only a fraction $\tau \in [0,1]$ of (x_i) while leaving a fraction $1 - \tau$ unchanged:

$$p_{cs}(i, y) = (1 - \tau)\delta_{x_i,y} + \tau P(y|X_i). \quad [4]$$

Here, $\delta_{x_i,y} = 1$ if $x_i = y$ and 0 otherwise. In principle, τ needs to be optimized depending on the evolutionary distance over which homologous sequences are to be found, in a similar way as the substitution matrix with optimum diversity might be chosen. In practice, we have found that, as with substitution matrices, a single diversity works well for the entire range of evolutionary distances (*SI Appendix*).

Fig. 1*B* illustrates the calculation of expected mutation probabilities $P(y|X_i)$ for a cysteine residue (highlighted in yellow) at position i belonging to a zinc-finger motif. Three profiles similar to the sequence window X_i (red box) are shown, whose central columns contribute to the context-specific sequence profile $p(i, y) = P(y|X_i)$ at position i with weights $P(p_k|X_i)$ of 7%, 60%, and 3%, respectively. With the resulting profile (*Lower*), a profile-to-sequence search can be performed, e.g., by using PSI-BLAST, which is equivalent to a sequence search with context-specific amino acid similarity scores (Eq. **1**). In this example, the context-specific scheme recognizes the sequence context of the cysteine and correctly assigns a zinc-finger profile a high weight, resulting in a highly conserved cysteine.

The context-specificity paradigm is not restricted to sequences but applies equally well to sequence profiles or profile hidden Markov models (HMMs) (*Materials and Methods*). It can therefore be used in profile-to-sequence (9, 32, 33) and profile-to-profile (8, 10, 34–37) comparison, for example.

Our method CS-BLAST for context-specific protein sequence searching is a simple extension of BLAST. First, a context-specific sequence profile is generated for the query sequence as described. This step is very fast. Then PSI-BLAST is jump-started with this profile. PSI-BLAST is extended to the context-specific case in an analogous way (CSI-BLAST) (*Materials and Methods*).

Benchmark

The homology detection performance of our context-specific method CS-BLAST and standard NCBI BLAST is evaluated on a benchmark dataset derived from SCOP 1.73 (38), filtered to a maximum pairwise sequence identity of 20% (SCOP20, 6,616 domains). SCOP is a database of protein domains with known structure, hierarchically ordered by class, fold, superfamily, and family. Following a standard procedure, we consider all domains from the same superfamily to be homologous (*true positives*) and all pairs from different SCOP folds to be nonhomologous (*false positives*). Domain pairs from the same fold but different superfamilies are ignored.

We randomly assign members of every fifth fold in SCOP20 to the *optimization set* (1,329 domains), the others to the *test set* (5,287 domains). By using the optimization set, we determined the best values for the pseudocount admixture ($\tau = 0.9$) and the window weights ($w_{\text{center}} = 1.6$, $\beta = 0.85$). The values for the window length ($l = 13$) and the context library size ($K = 4{,}000$) are a trade-off between sensitivity and time efficiency (see *SI Appendix*).

We perform an all-against-all comparison of the test-set domains and count the true and false positive hits at various E-value thresholds (Fig. 2*A*). To avoid a few large families from

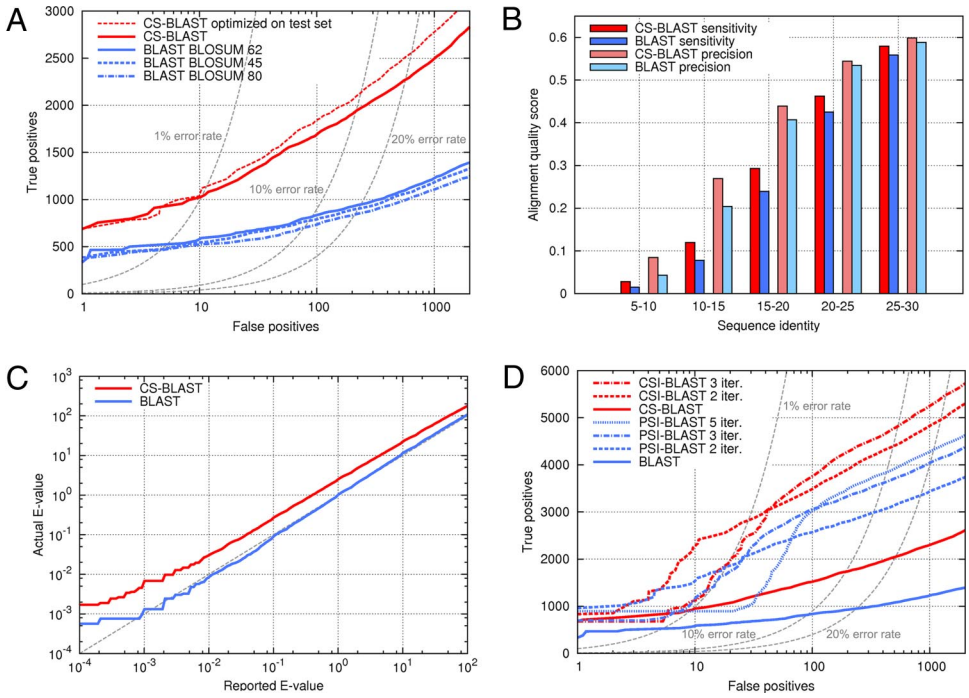

Fig. 2. Context information improves search performance and alignment quality. (*A*) Homology detection benchmark on SCOP20 dataset: true positives (pairs from the same SCOP superfamily) versus false positives (pairs from different folds). CS-BLAST detects 138% more true positives than BLAST at 10% error rate. (*B*) CS-BLAST has better average alignment sensitivity and precision than BLAST over the entire range of sequence identities of the aligned pairs. (*C*) Actual versus reported E-values on the SCOP20 dataset show that CS-BLAST E-values are too optimistic by a factor of 3 to 5. (*D*) Same benchmark as *A* (note different *y*-scales), but comparing CSI-BLAST with PSI-BLAST for one to five iterations. Two CSI-BLAST iterations are more sensitive than five PSI-BLAST iterations.

dominating the benchmark, we weight each true and false positive pair with 1/(size of SCOP family of first domain). Compared with NCBI BLAST (version 2.2.19, BLOSUM62, default parameters), CS-BLAST detects 139% more homologs at a cumulative error rate of 20%, 138% more at 10%, and, for the easiest cases at 1% error rate, still 96% more. To get an idea of the upside potential when parameters are trained on a larger set, we optimized w_{center}, β, and τ directly on the test set (red broken trace). These parameters ($w_{\text{center}} = 1.3$, $\beta = 0.9$, and $\tau = 0.95$) are used in the official version of CS-BLAST.

To assess the alignment quality, we compare predicted sequence alignments to gold-standard structural alignments generated by TM-align (39). We start by randomly picking up to 10 domain pairs from each family in SCOP 1.73, requiring a maximum sequence identity of 30%, and aligning each pair with TM-align. Those domains that are not well superposable (TM-align score < 0.6) are discarded. This results in 11,457 domain pairs from 5,747 different domains. With each of the 5747 domains we perform a CS-BLAST and NCBI BLAST search against a database consisting of all domains belonging to the same family as the query domain and evaluate the quality of the predicted alignments for those pairs with a structural reference alignment. The alignment quality is assessed by 2 standard performance measures: *Alignment sensitivity* is the fraction of structurally aligned residue pairs that are correctly predicted, i.e., pairs correctly aligned/pairs structurally alignable. *Alignment precision* is defined as the fraction of aligned residue pairs in the predicted alignment that are correct, i.e., pairs correctly aligned/pairs aligned. Fig. 2*B* plots alignment sensitivity and precision for various sequence identity bins. CS-BLAST is able to improve the BLAST results over the entire range of sequence identities, especially for the difficult alignments. Very similar results are obtained when reference alignments are generated with DALI (40).

Another critical aspect for database search tools is the reliability of the reported E-values. The E-value of a match is an estimate of the number of chance hits to be expected with a score better than that of the database match. We check the reliability of CS-BLAST E-values by using the all-against-all searches of Fig. 2*A*. We count the number of false positives at a given E-value threshold, which, together with the size of the benchmark database, allows us to derive the *actual* E-value. Fig. 2*C* plots the actual against the reported E-value. NCBI BLAST's reported E-values are nearly identical to the observed ones. CS-BLAST E-values are too optimistic by a factor of ≈3–5, e.g., a reported E-value of 10^{-3} corresponds to an E-value of 5×10^{-3}. Considering that this deviation is quite small and that it changes little with E-value, it should be easy to accommodate in practice.

Finally, we evaluate the homology detection performance of CSI-BLAST, the context-specific version of PSI-BLAST, on the benchmark of Fig. 2*A*. Because, typically, PSI-BLAST searches are done with a large sequence database, such as the nonredundant protein database (NR) at NCBI (41), to build diverse profiles, only the last search is performed against our benchmark database; all previous iterations use the full NR database (E-value inclusion thresholds set to 1×10^{-3} for PSI-BLAST and 2×10^{-4} for CSI-BLAST). Fig. 2*D* plots true positives versus false positives detected by PSI-BLAST and CSI-BLAST after up to 5 search iterations. (The trace for CSI-BLAST with 5 iterations has been omitted because it does not significantly improve over 3 iterations anymore. The traces for one iteration are the same as in Fig. 2*A*.) Remarkably, 2 iterations of CSI-BLAST are more sensitive than 5 rounds of standard PSI-BLAST (≈15% more homologs detected). This result surprised us. We had expected that context-specific profiles would only marginally improve sensitivity over standard sequence profiles, because profiles already contain family- and position-specific mutation rates. However, the lead of CS-BLAST over BLAST is even extended in absolute terms after the second iteration, demonstrating that the context profiles contain local information from *analogous* sequences (i.e., with similar

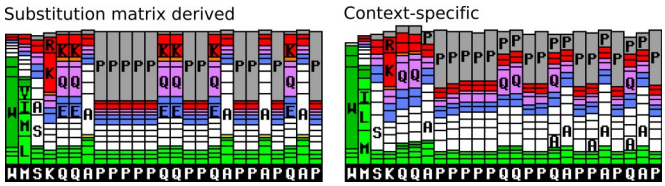

Fig. 3. Proline-rich region in human transcription factor SOX-9. The mutation profile computed with substitution matrix pseudocounts (*Left*) overestimates the conservation in this region. The context-specific profile (*Right*) shows weaker conservation of prolines, alanines, and glutamines, and increased presence of these residues in neighboring columns.

sequence context) that is partially independent of information from the *homologous* sequences in the profile.

Example: Activation Domain of SOX-9

Fig. 1*B* gave an example in which the context-specific method led to above-average conservation of Zn-finger cysteines. In practice, it will be equally important to be able to guess which residues are conserved less than average. As an example, Fig. 3 presents profiles of a region from the activation domain of human SOX-9 transcription factor, generated with substitution matrix pseudocounts (*Left*) and context-specific pseudocounts (*Right*). Because this region is natively disordered, its sequence is only very weakly conserved. The substitution matrix method assigns the same amino acid distribution to the prolines as it would to a proline in a globular domain. The context-specific method, however, mixes the pseudocounts mainly from contexts that are also disordered, weakly conserved, and have a similar, biased amino acid distribution. Therefore, its profile exhibits below-average conservation of prolines, alanines, and glutamines while having higher overall probabilities for these residues.

Discussion

Sequence context is much more powerful than a single residue in predicting which amino acids that particular residue is likely to mutate into (Fig. 2 *A* and *B*). Because this context information is as easy to get as the sequence itself, it is surprising that sequence context is practically never exploited. The main reason seems to be the focus of past research on structural context, with its limitation to proteins of known structures (17, 18, 20, 21, 27). Another reason may be the challenge to develop sequence context-specific methods that can compete with traditional context-free methods such as BLAST and PSI-BLAST in speed and usability (26, 27). We have shown how context-specific pseudocounts can be used in combination with existing profile-based methods to extend residue-centered sequence comparison to the context-specific case, without loss of speed or usability.

As examples, we have built context-specific versions of BLAST and PSI-BLAST that considerably improve their performance at very little runtime overhead. For a typical protein of length $L = 250$ and a library size of $K = 4{,}000$, the computation of the context-specific profile requires ≈1 s CPU time. Also, runtime scales favorably, $T \propto KlL$ (*SI Appendix*, Fig. S1). (Note that HMMSUM's runtime scales as $T \propto K^2$L, which places a strict practical limit on the number K of states/contexts in HMMSUM.) Because the output of CS-BLAST and CSI-BLAST is generated by the BLAST and PSI-BLAST programs themselves, users do not have to get accustomed to different command line options or output formats, and updates to the BLAST package will directly benefit the context-specific versions. The only caveat is that E-values need to be corrected by a factor of 3 to 5 (Fig. 2*C*). We expect CS-BLAST to be useful to find homologs for *singleton sequences*, because, for these, the lack of homologs precludes the use of profile-to-sequence search methods such as PSI-BLAST.

A pleasant surprise is the extent of improvements of sequence profiles through context-specific pseudocounts (Fig. 2*D*), even though profiles already contain evolutionary information on position- and family-specific mutation probabilities. Hence, the information from locally similar, *analogous* sequences that are contained in the context profiles is at least partly orthogonal to the evolutionary information in the *homologous* sequences that contribute to the sequence profiles. Consequently, we can expect improvements when applying the new paradigm to the pairwise comparison of sequence profiles (34–36) and profile HMMs (10, 37), or to hierarchical multiple sequence alignment programs (8, 42).

It is possible to extend Dirichlet mixture pseudocounts (29, 30) to the context-specific case. This would yield an alternative formulation of context-specific sequence comparison that is worth exploring. In that scheme, the context library would have K metaprofiles, i.e., multicolumn pseudocount priors. Each metaprofile would consist of l Dirichlet distributions and would be able to *emit* a profile with l columns. An advantage over the presented scheme might be that the diversity of each column in the metaprofiles is encoded by one additional parameter per column (the sum of all pseudocounts in a column), which might lead to better modeling of the profile contexts.

The paradigm presented here should be easily transferable to nucleotide sequences. The application to noncoding regions such as promoter regions and regions harboring putative noncoding RNAs (ncRNAs) is of particular interest. The low information content of nucleotide sequences and the often weak overall conservation in these regions render alignments between related species difficult, whereas reliable alignments offer enormous potential to identify functional regions (such as *cis*-regulatory elements or ncRNAs) through their interspecies conservation (see, e.g., ref. 43).

In summary, the paradigm of sequence context specificity offers greatly improved sensitivity and alignment quality in protein sequence comparison and is likely to hold similar advantages for nucleotide sequences. We believe that these advantages are sufficient to warrant a paradigm shift in biological sequence comparison, alignment, and molecular evolution from amino acid- and nucleotide-centric to context-specific methods.

Materials and Methods

Generalization to Sequence Profiles. To apply the paradigm to sequence profiles and profile HMMs, we show how to generalize the calculation of pseudocounts from the single sequence case in Eq. **2** to the case of sequence alignments, from which the profile is derived. In analogy to the sequence context X_i, we define the context of the query alignment at position i as $Q_i = (c_q(i - d,\cdot), \ldots, c_q(i + d,\cdot))$, where $c_q(j, x)$ are the counts of amino acid x at position j of the query alignment. These counts are obtained from the sequence profile $q(j, x)$ by multiplying with the effective number of sequences $N_q(j)$ at position j in the query alignment: $c_q(j, x) = N_q(j)\, q(j, x)$ (see *SI Appendix* for details). We now merely need to show how to generalize $P(X_i|p_k)$ to $P(Q_i|p_k)$, because all other transformations leading to Eq. **2** remain essentially unchanged. To derive $P(Q_i|p_k)$, we model the amino acid counts $c_q(i)$ with multinomial distributions. Because $N_q(j)$ can be real-valued, however, we replace the factorials in the multinomial distribution by Gamma functions ($n! = \Gamma(n + 1)$)

$$P(Q_i|p_k) = \prod_{j=-d}^{d} \left(\frac{\Gamma(N_q(i+j)+1)}{\Pi_{x=1}^{20}\Gamma(c_q(i+j,x)+1)} \prod_{x=1}^{20} p_k(j,x)^{c_q(i+j,x)} \right)^{w_j}. \quad [5]$$

Note that, because the factor containing the Gamma functions does not depend on k, it will cancel out during the normalization of $P(p_k|Q_i)$ (Eq. **2**, *SI Appendix*, Eq. **S9**). Similar to PSI-BLAST (9), we choose the pseudocount admixture τ in Eq. **3** depending on the diversity of the query alignment, $\tau = a(b + 1)/(b + N_q(i))$, where $a = 0.9$ and $b = 12.0$ have been determined on the training set as described in *SI Appendix*.

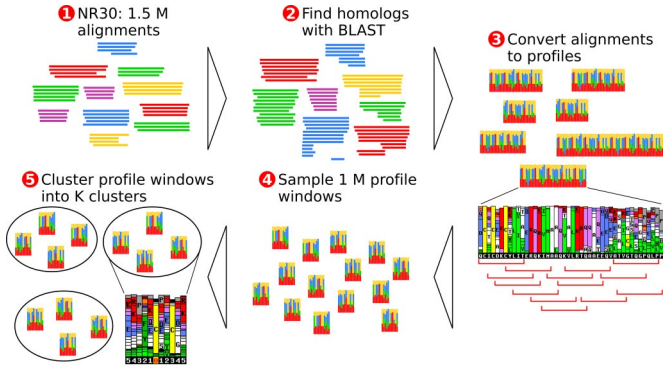

Fig. 4. Computation of the library of context profiles representing local sequence contexts. From a database (NR30) of 1.5M groups of aligned sequences covering the NR database, we select the 50,000 most diverse alignments and enrich these with homologs from a single BLAST search. The alignments are converted to sequence profiles and 1M profile windows are randomly sampled and used to train K context profiles (K = 500, 1,000, 2,000, 4,000) with the expectation maximization algorithm.

Generation of Context Profile Library. The quality of the predicted amino acid similarities depends to a large extent on the context profile library. The clustering procedure to derive this library is summarized in Fig. 4. We start with all sequences from the NR, clustered into groups with maximum intergroup sequence identity of 30% (NR30). In contrast to other approaches, in which only sequences with solved structure in the PDB were used (17, 18, 20, 21, 27), this guarantees an appropriate representation of all classes of local sequence contexts, such as membrane helices, natively unfolded regions, or highly repetitive sequences. From the 1.5 million cluster alignments in this NR30 database, we discard those with an effective number of sequences <2.5 (see *SI Appendix*) and jump-start a PSI-BLAST search against the full NR database with each of the remaining alignments (E-value threshold = 0.001). This ensures an alignment diversity that is sufficient to produce mutation probabilities in the same range as the BLOSUM62 matrix. After converting the alignments to profiles, we randomly sample 1 million training profile windows of length l from the full-length profile database. For a fixed number of context profiles (K = 500, 1,000, 2,000, 4,000) we determine the profile amino acid probabilities and the profile prior probabilities $P(p_k)$ by maximizing the total likelihood that the training profile windows are emitted by the context profiles. The maximization is done with the expectation maximization (EM) algorithm (44) (see *SI Appendix*).

Appendix: Availability of Datasets and Executables. The context profile library, all benchmark datasets, and results data files can be downloaded from ftp://toolkit.lmb.uni-muenchen.de/csblast. CS-BLAST executables for Linux (32 and 64 bit), Windows, and Mac are freely available for academic users. A free CS-BLAST webserver can be accessed at http://toolkit.lmb.uni-muenchen.de/cs_blast.

ACKNOWLEDGMENTS. We thank Michael Remmert for computational support and Nick Grishin and 2 anonymous referees for their very helpful comments. J.S. thanks Andrei Lupas for making his first 5 years in computational biology such an exciting and insightful experience.

1. Dayhoff M, Schwartz R, Orcutt B (1978) A model of evolutionary change in proteins. *Atlas Protein Sequence Struct* 5:345–352.
2. Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* 89:10915–10919.
3. Gonnet GH, Cohen MA, Benner SA (1992) Exhaustive matching of the entire protein sequence database. *Science* 256:1443–1445.
4. Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 48:443–453.
5. Smith TF, Waterman MS (1981) Identification of common molecular subsequences. *J Mol Biol* 147:195–197.
6. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410.
7. Pearson WR (1991) Searching protein sequence libraries: Comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics* 11:635–650.
8. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22:4673–4680.
9. Altschul SF, *et al.* (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402.
10. Söding J (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics* 21:951–960.
11. Henikoff JG, Henikoff S (1996) Using substitution probabilities to improve position-specific scoring matrices. *Comput Appl Biosci* 12:135–143.
12. Tatusov RL, Altschul SF, Koonin EV (1994) Detection of conserved segments in proteins: Iterative scanning of sequence databases with alignment blocks. *Proc Natl Acad Sci USA* 91:12091–12095.
13. Yu YK, Wootton JC, Altschul SF (2003) The compositional adjustment of amino acid substitution matrices. *Proc Natl Acad Sci USA* 100:15688–15693.
14. Jones DT, Taylor WR, Thornton JM (1994) A mutation data matrix for transmembrane proteins. *FEBS Lett* 339:269–275.
15. Ng PC, Henikoff JG, Henikoff S (2000) PHAT: A transmembrane-specific substitution matrix. *Bioinformatics* 16:760–766.
16. Mueller T, Rahmann S, Rehmsmeier M (2001) Non-symmetric score matrices and the detection of homologous transmembrane proteins. *Bioinformatics* 17:182–189.
17. Overington J, Donnelly D, Johnson MS, Sali A, Blundell TL (1992) Environment-specific amino acid substitution tables: Tertiary templates and prediction of protein folds. *Protein Sci* 1:216–226.
18. Rice DW, Eisenberg D (1997) A 3D–1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. *J Mol Biol* 267:1026–1038.
19. Shi J, Blundell TL, Mizuguchi K (2001) FUGUE: Sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J Mol Biol* 310:243–257.
20. Goonesekere NC, Lee B (2008) Context-specific amino acid substitution matrices and their use in the detection of protein homologs. *Proteins* 71:910–919.
21. Gelly JC, Chiche L, Gracy J (2005) EvDTree: Structure-dependent substitution profiles based on decision tree classification of 3D environments. *BMC Bioinformatics* 6:4.
22. Jung J, Lee B (2000) Use of residue pairs in protein sequence-sequence and sequence-structure alignments. *Protein Sci* 9:1576–1588.
23. Crooks GE, Green RE, Brenner SE (2005) Pairwise alignment incorporating dipeptide covariation. *Bioinformatics* 21:3704–3710.
24. Gambin A, Lasota S, Szklarczyk R, Tiuryn J, Tyszkiewicz J (2002) Contextual alignment of biological sequences (Extended abstract). *Bioinformatics* 18(Suppl 2):116–127.
25. Gambin A, Otto R (2005) Contextual multiple sequence alignment. *J Biomed Biotechnol* 2005:124–131.
26. Baussand J, Deremble C, Carbone A (2007) Periodic distributions of hydrophobic amino acids allows the definition of fundamental building blocks to align distantly related proteins. *Proteins* 67:695–708.
27. Huang YM, Bystroff C (2006) Improved pairwise alignments of proteins in the Twilight Zone using local structure predictions. *Bioinformatics* 22:413–422.
28. Han KF, Baker D (1996) Global properties of the mapping between local amino acid sequence and local structure in proteins. *Proc Natl Acad Sci USA* 93:5814–5818.
29. Sjoelander K, *et al.* (1996) Dirichlet mixtures: A method for improved detection of weak but significant protein sequence homology. *Comput Appl Biosci* 12:327–345.
30. Durbin R, Eddy S, Krogh A, Mitchison G (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids* (Cambridge Univ Press, Cambridge), pp 117–118.
31. Altschul SF (1991) Amino acid substitution matrices from an information theoretic perspective. *J Mol Biol* 219:555–565.
32. Gribskov M, McLachlan AD, Eisenberg D (1987) Profile analysis: Detection of distantly related proteins. *Proc Natl Acad Sci USA* 84:4355–4358.
33. Eddy SR (1998) Profile hidden Markov models. *Bioinformatics* 14:755–763.
34. Rychlewski L, Jaroszewski L, Li W, Godzik A (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci* 9:232–241.
35. Yona G, Levitt M (2002) Within the twilight zone: A sensitive profile-profile comparison tool based on information theory. *J Mol Biol* 315:1257–1275.
36. Sadreyev R, Grishin N (2003) COMPASS: A tool for comparison of multiple protein alignments with assessment of statistical significance. *J Mol Biol* 326:317–336.
37. Madera M (2008) Profile Comparer (PRC): A program for scoring and aligning profile hidden Markov models. *Bioinformatics* 24:2630–2631.
38. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 247:536–540.
39. Zhang Y, Skolnick J (2005) TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* 33:2302–2309.
40. Holm L, Sander C (1996) Mapping the protein universe. *Science* 273:595–603.
41. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2008) GenBank. *Nucleic Acids Res* 36:25–30.
42. Notredame (2007) Recent evolutions of multiple sequence alignment algorithms. *PLoS Comput Biol* 3:e123.
43. Stark, *et al.* (2007) Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature* 450:219–232.
44. Dempster A, Laird N, Rubin D (1977) Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc* 39:1–38.