Improved DNA-versus-Protein Homology Search for Protein Fossils

Yin Yao[1,2] and Martin C. Frith[2,1,3]

[1] Graduate School of Frontier Sciences, University of Tokyo
[2] Artificial Intelligence Research Center, AIST
[3] Computational Bio Big-Data Open Innovation Laboratory (CBBD-OIL), AIST

Abstract. Protein fossils, i.e. noncoding DNA descended from coding DNA, arise frequently from transposable elements (TEs), decayed genes, and viral integrations. They can reveal, and mislead about, evolutionary history and relationships. They have been detected by comparing DNA to protein sequences, but current methods are not optimized for this task. We describe a powerful DNA-protein homology search method. We use a 64×21 substitution matrix, which is fitted to sequence data, automatically learning the genetic code. We detect subtly homologous regions by considering alternative possible alignments between them, and calculate significance (probability of occurring by chance between random sequences). Our method detects TE protein fossils much more sensitively than `blastx`, and $> 10\times$ faster. Of the $\sim$7 major categories of eukaryotic TE, three have not been found in mammals: we find two of them in the human genome, polinton and DIRS/Ngaro. This method increases our power to find ancient fossils, and perhaps to detect non-standard genetic codes. The alternative-alignments and significance paradigm is not specific to DNA-protein comparison, and could benefit homology search generally.

1 Introduction

Genomes are littered with protein fossils, old and young. They can be found by comparing DNA to known proteins: new transposable element (TE) families have been discovered in this way [25]. An interesting class of protein fossils comes from ancient integrations of viral DNA into genomes, enabling the field of paleovirology [17]. The DNA sequences of protein fossils often have similarity to distantly-related genomes (e.g. mammal versus fish), simply because the parent gene evolved slowly, so it is important to know that they are protein fossils in order to understand this similarity [28]. DNA-protein homology search is also used to classify DNA reads from unknown microbes, including nanopore and PacBio reads with many sequencing errors [16]. DNA-protein comparison can be used to find frameshifts during evolution of functional proteins [26], and programmed ribosomal frameshifts [35]. A more specialized and complex kind of DNA-protein comparison, outside this study's scope, considers introns and other gene features to identify genes.

DNA-protein homology search is a classical problem with many old solutions [23, 30, 12, 15, 39, 13, 11, 20, 4, 22, 34]. A notable one is "three-frame alignment" [39], which we believe is the simplest and fastest reasonable way to do frameshifting DNA-protein alignment. Nevertheless, we can significantly improve DNA-protein homology search in these aspects:

- Better parameters for the (dis)favorability of substitutions, deletions, insertions, and frameshifts. Most previous methods use standard parameters such as the BLOSUM62 substitution matrix, which is designed for functional proteins, and likely completely inappropriate for protein fossils. We optimize these parameters by fitting them to sequence data.
- Instead of a 20×20 substitution matrix, use a 64×21 matrix (64 codons × 20 amino acids plus STOP). This allows e.g. preferred alignment of asparagine (which is encoded by `aac` and `aat`) to `agc` than to `tca`, which both encode serine.
- Incorporate frameshifts into affine gaps. Because gaps are somewhat rare but often long, it is standard to disfavor opening a gap more than extending a gap. However, most previous methods favor frameshifts equally whether isolated or contiguous with a longer gap.
- Detect homologous regions based on not just one alignment between them, but on many possible alternative alignments. This is expected to detect subtle homology more powerfully [1, 6].
- Calculate significance, i.e. the probability of such a strong similarity occurring by chance between random sequences. To this day, for ordinary alignment, BLAST can only calculate significance for a few hardcoded sets of substitution and gap parameters. We can do it for any parameters, for similarities based on many alternative alignments.

We also aimed for maximum simplicity and speed, inspired by three-frame alignment.

2 Methods

2.1 Alignment elements

We define a DNA-protein alignment to consist of: matches (3 bases aligned to 1 amino acid), base insertions, and base deletions. To keep things simple, insertions are not allowed between bases aligned to one amino acid. A deletion of length not divisible by 3 leaves "dangling" bases (Fig. 1): for simplicity, we do not attempt to align these (equivalently, align them to the amino acid with score 0).

2.2 Scoring scheme

An alignment's score is the sum of:

- Score for aligning amino acid x to base triplet Y: S_{xY}

```
Ser-TyrAlaThrMetLeuTrpAspGln--Leu***
tctCtat---acg--cctctga-atcagCAttctaa
```

Fig. 1. Example of a DNA-versus-protein alignment. `***` indicates a protein end from translation of a stop codon. Insertions are bold uppercase. "Dangling" bases, left by deletions of length not divisible by 3, are underlined gray.

- Score for an insertion of k bases: $a_I + b_I k + \begin{cases} 0 & \text{if } k \bmod 3 = 0 \\ f_I & \text{if } k \bmod 3 = 1 \\ g_I & \text{if } k \bmod 3 = 2 \end{cases}$
- Score for a deletion of k bases: $a_D + b_D k + \begin{cases} 0 & \text{if } k \bmod 3 = 0 \\ f_D & \text{if } k \bmod 3 = 1 \\ g_D & \text{if } k \bmod 3 = 2 \end{cases}$

This scheme extravagantly uses 4 frameshift parameters (f_I, g_I, f_D, g_D), because it's based on a probability model with 4 frameshift transitions (Fig. 2), and we can't think of a good way to simplify the model. Overall, our alignment scheme is similar to FramePlus [13] and especially to aln [11].

2.3 Finding a maximum-score local alignment

A basic approach is to find a maximum-score alignment between any parts of a protein sequence $R_0 \ldots R_{M-1}$ and a DNA sequence $q_0 \ldots q_{N-1}$. Let Q_j mean the triplet q_j, q_{j+1}, q_{j+2}. We can do these calculations for $0 \le i \le M$ and $0 \le j \le N$:

$$\begin{aligned}
y_1 &= Y_{i-1\ j-2} + [b_D + f_D] & z_1 &= Z_{i\ j-1} + [b_I + f_I] \\
y_2 &= Y_{i-1\ j-1} + [2b_D + g_D] & z_2 &= Z_{i\ j-2} + [2b_I + g_I] \\
y_3 &= Y_{i-1\ j} + [3b_D] & z_3 &= Z_{i\ j-3} + [3b_I] \\
X_{ij} &= \max(X_{i-1\ j-3} + S_{R_{i-1}Q_{j-3}},\ y_1,\ y_2,\ y_3,\ z_1,\ z_2,\ z_3,\ 0) \\
Y_{ij} &= \max(X_{ij} + a_D,\ y_3) & Z_{ij} &= \max(X_{ij} + a_I,\ z_3)
\end{aligned}$$

The boundary condition is: if $i < 0$ or $j < 0$, $X_{ij} = Y_{ij} = Z_{ij} = -\infty$. The maximum possible alignment score is $\max(X_{ij})$, and an alignment with this score can be found by a standard traceback [5].

For each (i, j) this algorithm retrieves 7 previous results, and performs 9 pairwise maximizations and 9 additions (which could be reduced to 6 additions if each insertion cost equals its corresponding deletion cost). This is slightly slower than three-frame alignment, which retrieves 5 previous results and performs 7 pairwise maximizations and 6 additions.

2.4 Probability model

The preceding algorithm is equivalent to finding a maximum-probability path generating the sequences, through a probability model (Fig. 2). The score and

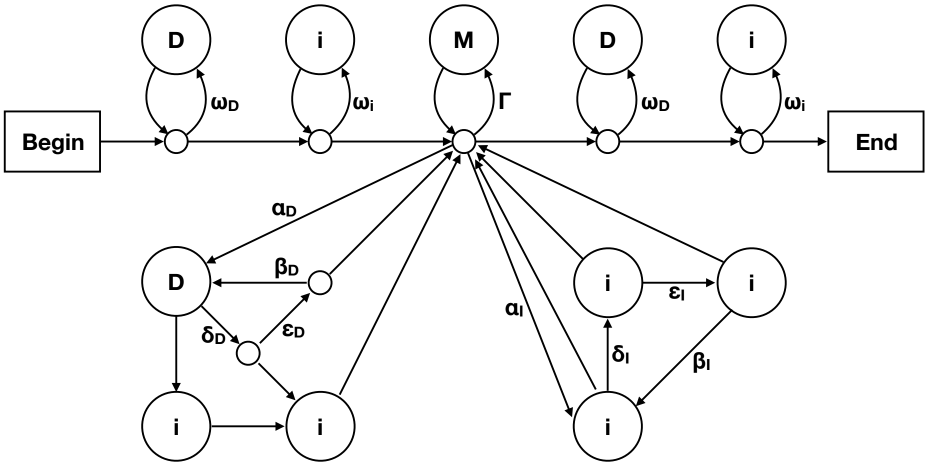

Fig. 2. A probability model for related DNA and protein sequences. The arrows are labeled with probabilities of traversing them. Each pass through an **i** state generates one base $y \in \{\texttt{a}, \texttt{c}, \texttt{g}, \texttt{t}\}$, with probabilities ψ_y. Each pass through a **D** state generates one amino acid x, with probabilities ϕ_x. Each pass through the **M** state generates one amino acid x aligned to three bases $Y = y_1 y_2 y_3$, with probabilities π_{xY}. The two bottom-left **i** states correspond to "dangling" bases.

model parameters are related like this:

$$S'_{xY} = \exp\left(\frac{S_{xY}}{t}\right) = \frac{\mathit{\Gamma}}{\omega_D \omega_i^3} \cdot \frac{\pi_{xY}}{\phi_x \psi_Y}$$

$$a'_I = \exp\left(\frac{a_I}{t}\right) = \frac{\alpha_I(1-\beta_I)}{\beta_I} \qquad a'_D = \exp\left(\frac{a_D}{t}\right) = \frac{\alpha_D(1-\beta_D)}{\beta_D}$$

$$b'_I = \exp\left(\frac{b_I}{t}\right) = \frac{\sqrt[3]{\beta_I \delta_I \epsilon_I}}{\omega_i} \qquad b'_D = \exp\left(\frac{b_D}{t}\right) = \sqrt[3]{\frac{\beta_D \delta_D \epsilon_D}{\omega_D}}$$

$$f'_I = \exp\left(\frac{f_I}{t}\right) = \frac{1-\delta_I}{1-\beta_I}\sqrt[3]{\frac{\beta_I^2}{\delta_I \epsilon_I}} \qquad f'_D = \exp\left(\frac{f_D}{t}\right) = \frac{1-\delta_D}{1-\beta_D}\sqrt[3]{\frac{\beta_D^2}{\delta_D \epsilon_D \omega_D^2}}\Big/\omega_i^2$$

$$g'_I = \exp\left(\frac{g_I}{t}\right) = \frac{1-\epsilon_I}{1-\beta_I}\sqrt[3]{\frac{\beta_I \delta_I}{\epsilon_I^2}} \qquad g'_D = \exp\left(\frac{g_D}{t}\right) = \frac{1-\epsilon_D}{1-\beta_D}\sqrt[3]{\frac{\beta_D \delta_D}{\epsilon_D^2 \omega_D}}\Big/\omega_i$$

Here ψ_Y is defined to be $\psi_{y_1}\psi_{y_2}\psi_{y_3}$, and t is an arbitrary positive constant (because multiplying all the score parameters by a constant makes no difference to alignment). An alignment score is then: $t \ln$[prob(path & sequences)/ prob(null path & sequences)], where a "null path" is a path that never traverses the $\mathit{\Gamma}$, α_D, or α_I arrows [10].

Balanced length probability A fundamental property of local alignment models is whether they are biased towards longer or shorter alignments [10]. If

ω_D and ω_i are large (close to 1) and $\mathit{\Gamma} + \alpha_D + \alpha_I$ is small, there is a bias in favor of shorter alignments. In the converse situation, there is a bias towards longer alignments. It can be shown (using the method of [10]) that our DNA-protein model is unbiased when

$$\frac{\mathit{\Gamma}}{\omega_D \omega_i^3} + \frac{a'_I b'_I (f'_I + g'_I b'_I + b'^2_I)}{1 - b'^3_I} + \frac{a'_D b'_D (f'_D + g'_D b'_D + b'^2_D)}{1 - b'^3_D} = 1\,. \tag{1}$$

2.5 Sum over all alignments passing through (i, j)

To find subtly homologous regions, we should assess their homology without fixing an alignment [1, 6]. In other words, we should use a homology score like this: $t \ln \left[\sum_{\text{paths}} \text{prob(path \& sequences)/prob(null path \& sequences)} \right]$. However, if the sum is taken over all possible paths, we learn nothing about location of the homologous regions, which is important if e.g. the DNA sequence is a chromosome. There is a kind of uncertainty principle here: the more we pin down the alignment, the less power we have to detect homology. As a compromise, we sum over all paths passing through one (protein, DNA) coordinate pair (i, j). This has two further benefits: it is approximated by the seed-and-extend search used for big sequence data, and we can calculate significance.

To calculate this sum over paths, we first run a Forward algorithm for $0 \le i \le M$ and $0 \le j \le N$:

$$\begin{aligned}
y_1 &= [b'_D f'_D] Y^F_{i-1\ j-2} & y_2 &= [b'^2_D g'_D] Y^F_{i-1\ j-1} & y_3 &= [b'^3_D] Y^F_{i-1\ j} \\
z_1 &= [b'_I f'_I] Z^F_{i\ j-1} & z_2 &= [b'^2_I g'_I] Z^F_{i\ j-2} & z_3 &= [b'^3_I] Z^F_{i\ j-3} \\
X^F_{ij} &= S'_{R_{i-1} Q_{j-3}} X^F_{i-1\ j-3} + y_1 + y_2 + y_3 + z_1 + z_2 + z_3 + 1 \\
Y^F_{ij} &= a'_D X^F_{ij} + y_3 & Z^F_{ij} &= a'_I X^F_{ij} + z_3
\end{aligned}$$

The boundary condition is: if $i < 0$ or $j < 0$, $X^F_{ij} = Y^F_{ij} = Z^F_{ij} = 0$. We then run a Backward algorithm for $M \ge i \ge 0$ and $N \ge j \ge 0$:

$$\begin{aligned}
y_1 &= [b'_D f'_D] Y^B_{i+1\ j+2} & y_2 &= [b'^2_D g'_D] Y^B_{i+1\ j+1} & y_3 &= [b'^3_D] Y^B_{i+1\ j} \\
z_1 &= [b'_I f'_I] Z^B_{i\ j+1} & z_2 &= [b'^2_I g'_I] Z^B_{i\ j+2} & z_3 &= [b'^3_I] Z^B_{i\ j+3} \\
X^B_{ij} &= S'_{R_i Q_j} X^B_{i+1\ j+3} + y_1 + y_2 + y_3 + z_1 + z_2 + z_3 + 1 \\
Y^B_{ij} &= a'_D X^B_{ij} + y_3 & Z^B_{ij} &= a'_I X^B_{ij} + z_3
\end{aligned}$$

The boundary condition is: if $i > M$ or $j > N$, $X^B_{ij} = Y^B_{ij} = Z^B_{ij} = 0$. Finally, $t \ln[X^F_{ij} X^B_{ij}]$ is the desired homology score, for all paths passing through (i, j).

2.6 Significance calculation

The just-described homology score is similar to that of "hybrid alignment", which has a conjecture regarding significance [37]. (Hybrid alignment sums over paths ending at (i, j), instead of passing through (i, j).) We make a similar

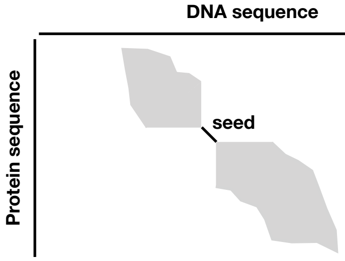

Fig. 3. Sketch of seed-and-extend heuristic for homology search.

conjecture. Suppose we compare a random i.i.d. protein sequence of length M and letter probabilities Φ_x to a random i.i.d. DNA sequence of length N and triplet probabilities Ψ_Y. We conjecture that the score $s_{\max} = t \ln[\max_{ij}(X^F_{ij} X^B_{ij})]$ follows a Gumbel distribution:

$$\mathrm{prob}(s_{\max} < s) = \exp(-KMNe^{-s/t})\,, \tag{2}$$

in the limit that M and N are large, provided that:

$$\Big(\sum_{x,Y} \Phi_x \Psi_Y S'_{xY}\Big) + \frac{a'_I b'_I (f'_I + g'_I b'_I + b'^2_I)}{1 - b'^3_I} + \frac{a'_D b'_D (f'_D + g'_D b'_D + b'^2_D)}{1 - b'^3_D} = 1\,. \tag{3}$$

Equation 3 is analogous to Equation 27 or 28 in [37], see also [10]. In practice, we assume that $\Phi_x = \phi_x$ and $\Psi_Y = \psi_Y$, which makes Equation 3 equivalent to Equation 1.

This conjecture leaves one unknown Gumbel parameter K. We estimate it by brute-force simulation of 50 pseudorandom sequence pairs [38], with $\Phi_x = \phi_x$, $\Psi_Y = \psi_Y$, $M = 200$ and $N = 602$, which takes zero human-perceptible run time.

2.7 Seed-and-extend heuristic

To find homologous regions in big sequence data, we use a BLAST-like seed-and-extend heuristic (Fig. 3) [2]. We first find "seeds": we currently use exact-matches (via the genetic code), which can be sensitive if short, but we could likely get better sensitivity per run time with inexact seeds [27,31]. Our seeds have variable length: starting from each DNA base, we get the shortest seed that occurs $\leq m$ times in the protein data [19]. We then try a gapless X-drop extension in both directions, and if the score achieves a threshold d, we try a "Forward" extension in both directions.

We use our Forward algorithm, modified for semi-global instead of local alignment. In each direction, we sum over alignments starting at the seed and ending anywhere: thus the algorithm's $+1$ is done only at the first (i, j) next to the seed, and we accumulate the sum $W = \sum_{ij} X^F_{ij}$. We run this algorithm in increasing order of antidiagonal $(3i + j)$ on the seed's right side (decreasing order on the left side). If X_{ij}^F is less than a fraction f of W accumulated over previous antidiagonals, we stop extending, which defines the boundary of the gray region in Fig. 3. The final homology score is $t\ln[W_{\text{left}}] + \text{seed score} + t\ln[W_{\text{right}}]$.

Sum-of-path algorithms are prone to numerical overflow [5]. To prevent that: once per 32 antidiagonals, we multiply all the X^F, Y^F, and Z^F values in the last six antidiagonals by a scaling factor of $1/W$.

A score with no alignment is disconcerting, so we get a representative alignment by a similar semi-global modification of our maximum-score alignment algorithm. To avoid redundancy, we prioritize homology predictions by score (breaking ties arbitrarily), and discard any prediction whose representative alignment shares an an (i, j) left or right end with a higher-priority prediction.

2.8 Fitting substitution & gap parameters to sequence data

We can fit the parameters to some related (unaligned) DNA and proteins, by an iterative Baum-Welch algorithm [5]. We implemented two versions of this: an exact $O(MN)$ version, and a seed-and-extend version. The seed-and-extend version, at each iteration, finds significantly homologous regions (with `-K1` filtering, see below) and gets expected counts from the seeds and extend regions (gray areas in Fig. 3). It does not infer ϕ_x, ψ_y, ω_D, or ω_i in the usual way: at each iteration, it sets $\phi_x = \sum_Y \pi_{xY}$, $\psi_y = \sum_{xij}(\pi_{x\,yij} + \pi_{x\,iyj} + \pi_{x\,ijy})/3$, and $\omega_D = \omega_i^3 =$ the value that satisfies Equation 1 (found by bisection with bounds $1 > \omega_i^3 > \beta_I\delta_I\epsilon_I$ and $1 > \omega_D > \beta_D\delta_D\epsilon_D$). We set $t = 3/\ln[2]$ to get scores in third-bit units.

3 Results

3.1 Parameter fitting

We applied our $O(MN)$ fitting to a set of human processed pseudogenes and their parent proteins from Pseudofam [21]. To avoid bias, we began the first iteration with $\pi_{xY} = 1/(21 \cdot 64)$. The fitting discovered the genetic code: for each codon Y, its encoded amino acid has maximum S_{xY}.

Sometimes, our fitting had an undesirable feature: the S_{xY} values for some `cg`-containing codons were all negative. This is presumably due to the well-known depletion of `cg` in human DNA, which can be captured in π_{xY} but not ψ_y. As an ad hoc fix, we set $\psi_Y = \sum_x \pi_{xY}$ (after $O(MN)$ fitting, and at each iteration of seed-and-extend fitting).

Next, we applied our seed-and-extend fitting to human chromosome 21 (hg38 chr21) and transposable element (TE) proteins from RepeatMasker 4.1.0 [29]. The result primarily favors genetic-code matches (Figure 4), and secondarily favours single `a`↔`g` or `c`↔`t` mismatches, e.g. asparagine scores +5 with `agc` and −14 with `tca`. The gap scores are $a_I, b_I, f_I, g_I = -28, -1, +3, 0$ and $a_D, b_D, f_D, g_D = -23, -1, +3, 0$. So frameshifts are not disfavored, perhaps because RepeatMasker's proteins are close to the fossils's most recent active ancestors. The positive $f_{I,D}$ values might be caused by the gap-length distribution not fitting the simple affine model, with an excess of length-1 and length-4 gaps [33].

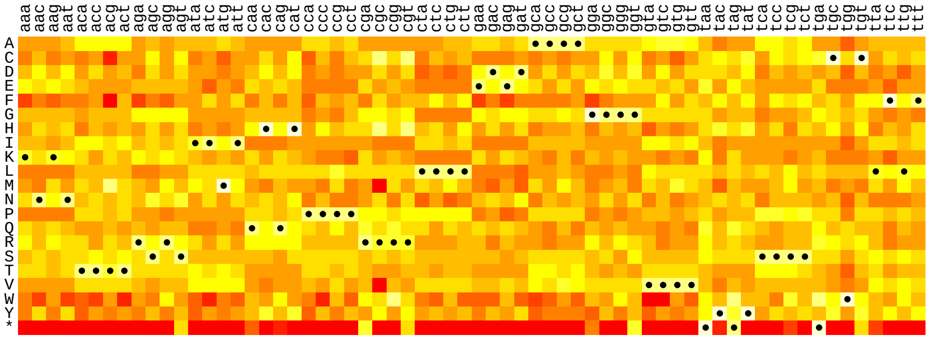

Fig. 4. Substitution matrix inferred from human chromosome 21 versus RepeatMasker proteins. Darker red means more disfavored and paler yellow means more favored. Black dots indicate the standard genetic code.

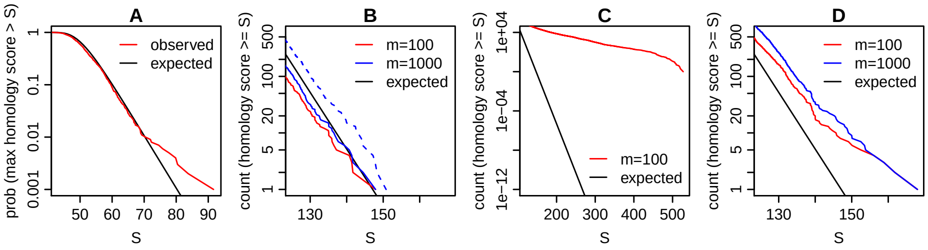

Fig. 5. Distributions of homology scores using the chr21-TE parameters. (**A**) Exact homology scores for random protein & codon sequences. (**B**) Seed-and-extend homology scores for random protein & DNA sequences. The dashed line shows a test with $(\mathit{\Phi}, \mathit{\Psi}) \neq (\phi, \psi)$. (**C**) Seed-and-extend homology scores for TE proteins & reversed chr21 without and (**D**) with simple-sequence masking.

3.2 Significance calculation & simple sequences

To test the accuracy of our significance estimates, for the chr21-TE parameters, we calculated $s_{\max}$ by our full Forward-Backward algorithm for 1000 pairs of random i.i.d. protein and codon sequences, with $\mathit{\Phi}_x = \phi_x$, $\mathit{\Psi}_Y = \psi_Y$, $M = 200$, and $N = 602$. The observed distribution of $s_{\max}$ agrees reasonably well with that predicted by Equation 2 (Fig. 5A).

To test whether our significance estimates apply to our seed-and-extend homology search, we compared one pair of random i.i.d. protein and DNA sequences, with $\mathit{\Phi}_x = \phi_x$, $\mathit{\Psi}_y = \psi_y$, and lengths equal to the number of unambiguous letters in the TE proteins and chr21. The search sensitivity depends on the seed parameter m: as m increases, sensitivity increases, and the distribution of homology scores approaches the Gumbel prediction (Fig. 5B).

We then considered $(\mathit{\Phi}, \mathit{\Psi}) \neq (\phi, \psi)$, because the marginal frequencies of π_{xY} differ from the letter abundances in the TE proteins and chr21, e.g. $\psi_{\mathtt{a}}$:$\psi_{\mathtt{c}}$:$\psi_{\mathtt{g}}$:$\psi_{\mathtt{t}}$ = 40:19:18:23 but chr21 is 29:21:21:29. So we compared another pair of random i.i.d. protein and DNA sequences, with Φ_x and Ψ_y equal to the frequencies in the TE proteins and chr21. In this test, the E-values (expected counts) were too low by a factor of about 3 (Fig. 5B).

Homology search is confounded by "simple sequences", e.g. `ttttctttttttcctt`, which evolve frequently and independently. There are various methods to suppress such false homologies, but most do not fully succeed [9, 8]. To illustrate, we compared reversed (but not complemented) chr21 to the TE proteins: this test has no true homologies, but we found many highly-significant homology scores (Fig. 5C). Our solution is to mask the DNA and protein with `tantan` [9], which eliminates extremely-significant false homologies, at least in this test (Fig. 5D). Further testing is warranted, e.g. here we used a default `tantan` parameter $r = 0.005$, but 0.02 was suggested for DNA-protein comparison [9].

3.3 Comparison to blastx

To test whether our homology search is more sensitive than standard methods, we compared chr21 to the TE proteins with NCBI BLAST 2.11.0:

```
makeblastdb -in RepeatPeps.lib -dbtype prot -out DB
blastx -query chr21.fa -db DB -evalue 0.1 -outfmt 7 > out
```

We repeated this comparison with our method (in LAST version 1177):

```
lastdb -q -c -R01 myDB RepeatPeps.lib
last-train --codon -X1 myDB chr21.fa > train.out
lastal -p train.out -D1e9 -m100 -K1 myDB chr21.fa > out
```

Option `-q` appends `*` to each protein; `-R01` lowercases simple sequence with `tantan`; `-c` requests masking of lowercase; `-X1` treats matches to unknown residues (which are frequent in these proteins) as neutral instead of disfavored; `-D1e9` sets the significance threshold to 1 random hit per 10^9 basepairs; `-m100` sets $m = 100$; `-K1` omits alignments whose DNA range lies in that of a higher-scoring alignment.

This test indicated that our method has much better sensitivity and speed. The single-threaded runtimes were 193 min for `blastx` and 18 min for `lastal`. `blastx` found alignments at 2604 non-overlapping sites on the two strands of chr21, of which all but 23 overlapped LAST alignments. LAST found alignments at 6640 non-overlapping sites, of which 4499 did not overlap `blastx` alignments. All but 21 of LAST's sites overlapped same-strand annotations by RepeatMasker open-4.0.6 - Dfam 2.0 (excluding Simple_repeat and Low_complexity) [29, 32], suggesting they are not spurious.

3.4 Discovery of missing TE orders in the human genome

Eukaryotic TEs have immense diversity, but can be classified into ~7 major orders: LTR, LINE, and tyrosine-recombinase (YR) retrotransposons, and DDE transposons, cryptons, helitrons, and polintons [36]. Three of them (YR retrotransposons, cryptons, polintons) have not been found in mammals [24, 3].

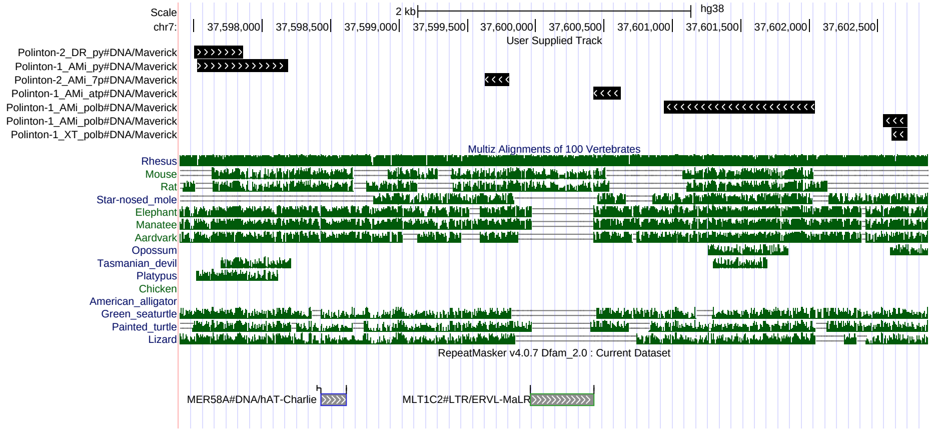

Fig. 6. An ancient polinton in human chromosome 7. Black bars: alignments to polinton proteins, arrows indicate +/− strand. Green: alignments to other vertebrate genomes [14]. Screen shot from `http://genome.ucsc.edu` [18].

By comparing the whole human genome (hg38) to RepeatMasker's TE proteins, we found two of these missing TE orders: YR retrotransposons and polintons. We found polinton alignments at 18 non-overlapping genome sites, with E-values as low as 1.8e-36. Five of these sites are clustered in chromosome 7, indicating that an ancient polinton was fragmented by insertion of an LTR element and an inversion (Fig. 6). We found both major superfamilies of YR retrotransposon: DIRS at 20 non-overlapping sites with min E-value 3e-45, and Ngaro at 4 non-overlapping sites with min E-value 5.1e-14.

4 Discussion

Our DNA-protein homology search method seems to be fast, specific, and highly sensitive. It should enable discovery of more ancient and subtle fossils, such as the human polinton, DIRS and Ngaro elements found here. So almost all known major TE categories have left traces in the human genome, suggesting an ability to spread broadly among eukaryotes.

Possible future improvements include better seeding, and using position-specific information on variability of a sequence family [5, 38]. Our significance calculation becomes inaccurate for short sequences, so a finite size correction would be useful [37]. Our method's parameter-fitting makes it versatile, but it would be better to use different parameters for fossils of different ages.

The sum-of-paths and significance paradigm is not specific to DNA-protein comparison, so could benefit homology search generally. A previous study made similar conjectures on significance of probabilistic homology scores [7]. We suspect those conjectures may be too broad: e.g. one set of substitution and gap scores corresponds to a range of probability models with different values of t [10], but only one t can appear in the Gumbel formula (Equation 2).

Acknowledgments We thank the Frith and Asai lab members for discussions that clarified our thinking.

References

1. Allison, L., Wallace, C.S., Yee, C.N.: Finite-state models in the alignment of macromolecules. J. Mol. Evol. **35**(1), 77–89 (Jul 1992)
2. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research **25**(17), 3389–3402 (1997)
3. Campbell, S., Aswad, A., Katzourakis, A.: Disentangling the origins of virophages and polintons. Current opinion in virology **25**, 59–65 (2017)
4. Csűrös, M., Miklós, I.: Statistical alignment of retropseudogenes and their functional paralogs. Molecular biology and evolution **22**(12), 2457–2471 (2005)
5. Durbin, R., Eddy, S., Krogh, A., Mitchison, G.: Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge University Press (1998)
6. Eddy, S.R.: A new generation of homology search tools based on probabilistic inference. Genome Inform **23**(1), 205–211 (Oct 2009)
7. Eddy, S.R.: A probabilistic model of local sequence alignment that simplifies statistical significance estimation. PLoS Comput Biol **4**(5), e1000069 (2008)
8. Frith, M.C.: Gentle masking of low-complexity sequences improves homology search. PLoS One **6**(12), e28819 (2011)
9. Frith, M.C.: A new repeat-masking method enables specific detection of homologous sequences. Nucleic acids research **39**(4), e23–e23 (2011)
10. Frith, M.C.: How sequence alignment scores correspond to probability models. Bioinformatics **36**(2), 408–415 (2020)
11. Gotoh, O.: Homology-based gene structure prediction: simplified matching algorithm using a translated codon (tron) and improved accuracy by allowing for long gaps. Bioinformatics **16**(3), 190–202 (Mar 2000)
12. Guan, X., Uberbacher, E.C.: Alignments of DNA and protein sequences containing frameshift errors. Comput. Appl. Biosci. **12**(1), 31–40 (Feb 1996)
13. Halperin, E., Faigler, S., Gill-More, R.: FramePlus: aligning DNA to protein sequences. Bioinformatics **15**(11), 867–873 (Nov 1999)
14. Harris, R.S.: Improved pairwise alignment of genomic DNA. Ph.D. thesis, The Pennsylvania State University (2007)
15. Huang, X., Zhang, J.: Methods for comparing a DNA sequence with a protein sequence. Bioinformatics **12**(6), 497–506 (1996)
16. Huson, D.H., Albrecht, B., Bağcı, C., Bessarab, I., Gorska, A., Jolic, D., Williams, R.B.: MEGAN-LR: new algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs. Biology direct **13**(1), 6 (2018)
17. Katzourakis, A., Gifford, R.J.: Endogenous viral elements in animal genomes. PLoS Genet. **6**(11), e1001191 (Nov 2010)
18. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., Haussler, D.: The human genome browser at UCSC. Genome research **12**(6), 996–1006 (2002)

19. Kiełbasa, S.M., Wan, R., Sato, K., Horton, P., Frith, M.C.: Adaptive seeds tame genomic sequence comparison. Genome research **21**(3), 487–493 (2011)
20. Ko, P., Narayanan, M., Kalyanaraman, A., Aluru, S.: Space-conserving optimal DNA-protein alignment. In: Proceedings. 2004 IEEE Computational Systems Bioinformatics Conference, 2004. CSB 2004. pp. 80–88. IEEE (2004)
21. Lam, H.Y., Khurana, E., Fang, G., Cayting, P., Carriero, N., Cheung, K.H., Gerstein, M.B.: Pseudofam: the pseudogene families database. Nucleic acids research **37**(suppl_1), D738–D743 (2009)
22. Lysholm, F.: Highly improved homopolymer aware nucleotide-protein alignments with 454 data. BMC Bioinformatics **13**(1), 230 (2012)
23. Peltola, H., Söderlund, H., Ukkonen, E.: Algorithms for the search of amino acid patterns in nucleic acid sequences. Nucleic acids research **14**(1), 99–107 (1986)
24. Poulter, R.T., Butler, M.I.: Tyrosine recombinase retrotransposons and transposons. Mobile DNA III pp. 1271–1291 (2015)
25. Pritham, E.J., Feschotte, C.: Massive amplification of rolling-circle transposons in the lineage of the bat *Myotis lucifugus*. Proceedings of the National Academy of Sciences **104**(6), 1895–1900 (2007)
26. Raes, J., Van de Peer, Y.: Functional divergence of proteins through frameshift mutations. Trends in Genetics **21**(8), 428–431 (2005)
27. Roytberg, M., Gambin, A., Noé, L., Lasota, S., Furletova, E., Szczurek, E., Kucherov, G.: On subset seeds for protein alignment. IEEE/ACM Transactions on Computational Biology and Bioinformatics **6**(3), 483–494 (2009)
28. Sheetlin, S.L., Park, Y., Frith, M.C., Spouge, J.L.: Frameshift alignment: statistics and post-genomic applications. Bioinformatics **30**(24), 3575–3582 (2014)
29. Smit, A., Hubley, R., Green, P.: RepeatMasker open-4.0. `http://www.repeatmasker.org` (2013–2015)
30. States, D., Botstein, D.: Molecular sequence accuracy and the analysis of protein coding regions. Proceedings of the National Academy of Sciences of the United States of America **88**(13), 5518 (1991)
31. Steinegger, M., Söding, J.: MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature biotechnology **35**(11), 1026–1028 (2017)
32. Storer, J., Hubley, R., Rosen, J., Wheeler, T.J., Smit, A.F.: The Dfam community resource of transposable element families, sequence models, and genome annotations. Mobile DNA **12**(1), 1–14 (2021)
33. Tanay, A., Siggia, E.D.: Sequence context affects the rate of short insertions and deletions in flies and primates. Genome biology **9**(2), R37 (2008)
34. Tzou, P.L., Huang, X., Shafer, R.W.: NucAmino: a nucleotide to amino acid alignment optimized for virus gene sequences. BMC bioinformatics **18**(1), 138 (2017)
35. Wang, R., Xiong, J., Wang, W., Miao, W., Liang, A.: High frequency of +1 programmed ribosomal frameshifting in *Euplotes octocarinatus*. Scientific Reports **6**, 21139 (2016)
36. Wells, J.N., Feschotte, C.: A field guide to eukaryotic transposable elements. Annual Review of Genetics **54**, 539–561 (2020)
37. Yu, Y.K., Hwa, T.: Statistical significance of probabilistic sequence alignment and related local hidden Markov models. J. Comput. Biol. **8**(3), 249–282 (2001)
38. Yu, Y.K., Bundschuh, R., Hwa, T.: Hybrid alignment: high-performance with universal statistics. Bioinformatics **18**(6), 864–872 (2002)
39. Zhang, Z., Pearson, W.R., Miller, W.: Aligning a DNA sequence with a protein sequence. J. Comput. Biol. **4**(3), 339–349 (1997)