

Contextual Alignment of Biological Sequences



> Radek Szklarczyk

>

> joint work with Ania Gambin, Sławomir Lasota, \
Jerzy Tiuryn and Jerzy Tyszkiewicz

>

> Warsaw University

-



Why to Compare Sequences?

Find similar regions in sequence: they may define a domain

- Useful when dealing with unknown sequence

Derive evolutionary relationships

- Existence of common ancestor

Key Property of Contextual Alignment

Substitution $\mathbf{A} \rightarrow \mathbf{V}$ depends on the amino acids before and after the substituted one

Original seq: ... **L** **A** **R** ...



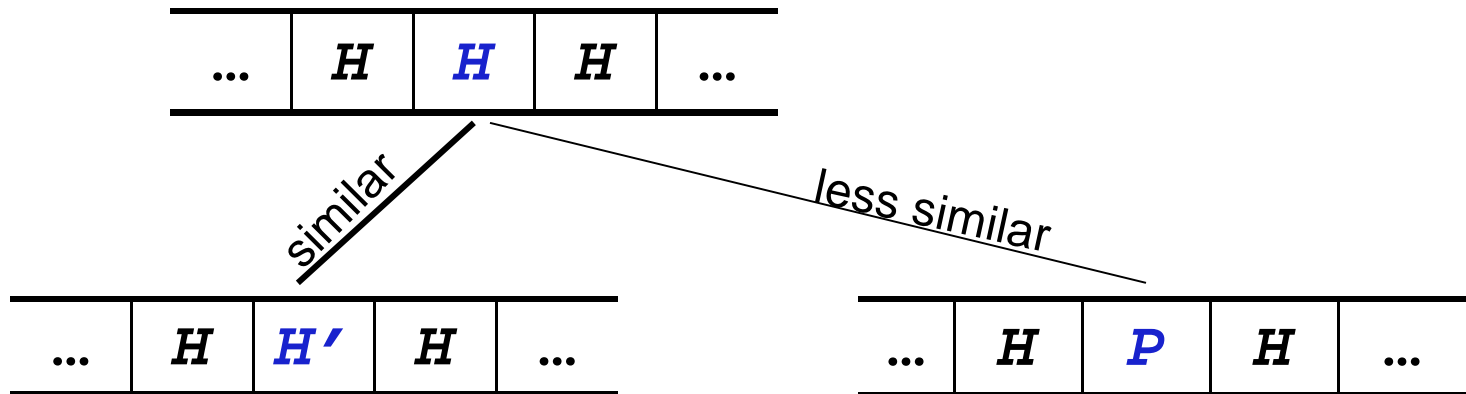
Mutated seq: ... **L** **V** **R** ...

$$\text{score}(S_{L, R}(\mathbf{A}, \mathbf{V})) = 3.2$$

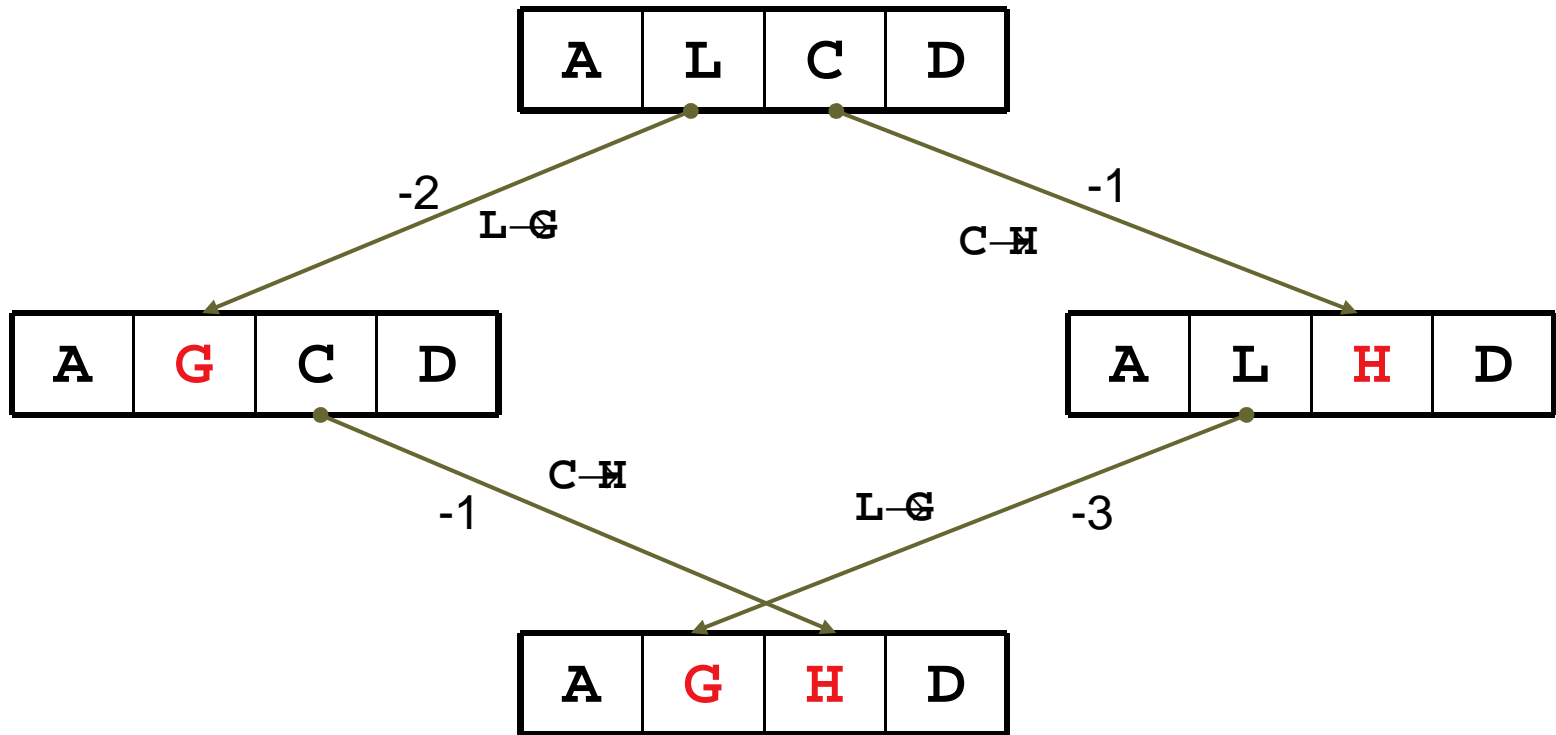
Insertions and deletions might have different score depending on surrounding amino acids

Why Contextual?

Proteins: sequence \Rightarrow structure \Rightarrow function



Order of operations matters



Note the different score for the same mutation **L → G** :

$$\text{score}(S_{A, C}(\mathbf{L}, \mathbf{G})) \neq \text{score}(S_{A, H}(\mathbf{L}, \mathbf{G}))$$

Example

aa no.	1	2	3	4	5	6	7
seq 1	E	A	-	-	C	G	T
seq 2	F	A	C	D	H	-	V

Three kinds of operations:

Substitution: e.g., $S_{E,H}(A,A)$, $S_{A,V}(C,H)$, $S(E,F)$, $S(T,V)$

Insertion: I_3

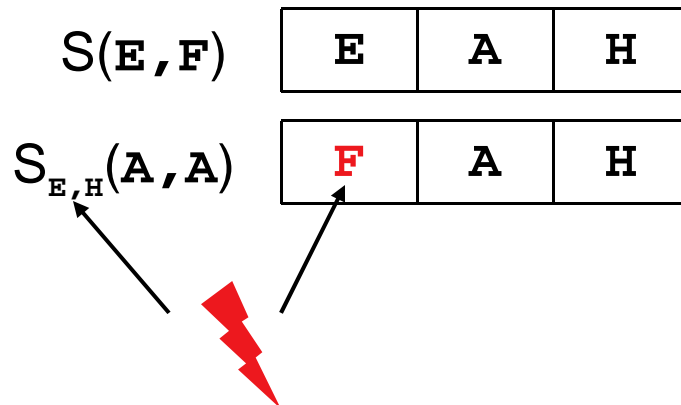
Deletion: D_6

An Example of Invalid Order

1	2	3	4	5	6	7
E	A	-	-	C	G	T
F	A	C	D	H	-	V

Let's consider two operations: substitution on position 1: $S(\mathbf{E}, \mathbf{F})$ and position 2: $S_{\mathbf{E}, \mathbf{H}}(\mathbf{A}, \mathbf{A})$.

Q: Is sequence $S(\mathbf{E}, \mathbf{F})$ followed by $S_{\mathbf{E}, \mathbf{H}}(\mathbf{A}, \mathbf{A})$ valid?



The only valid order is $S_{\mathbf{E}, \mathbf{H}}(\mathbf{A}, \mathbf{A}); S(\mathbf{E}, \mathbf{F})$

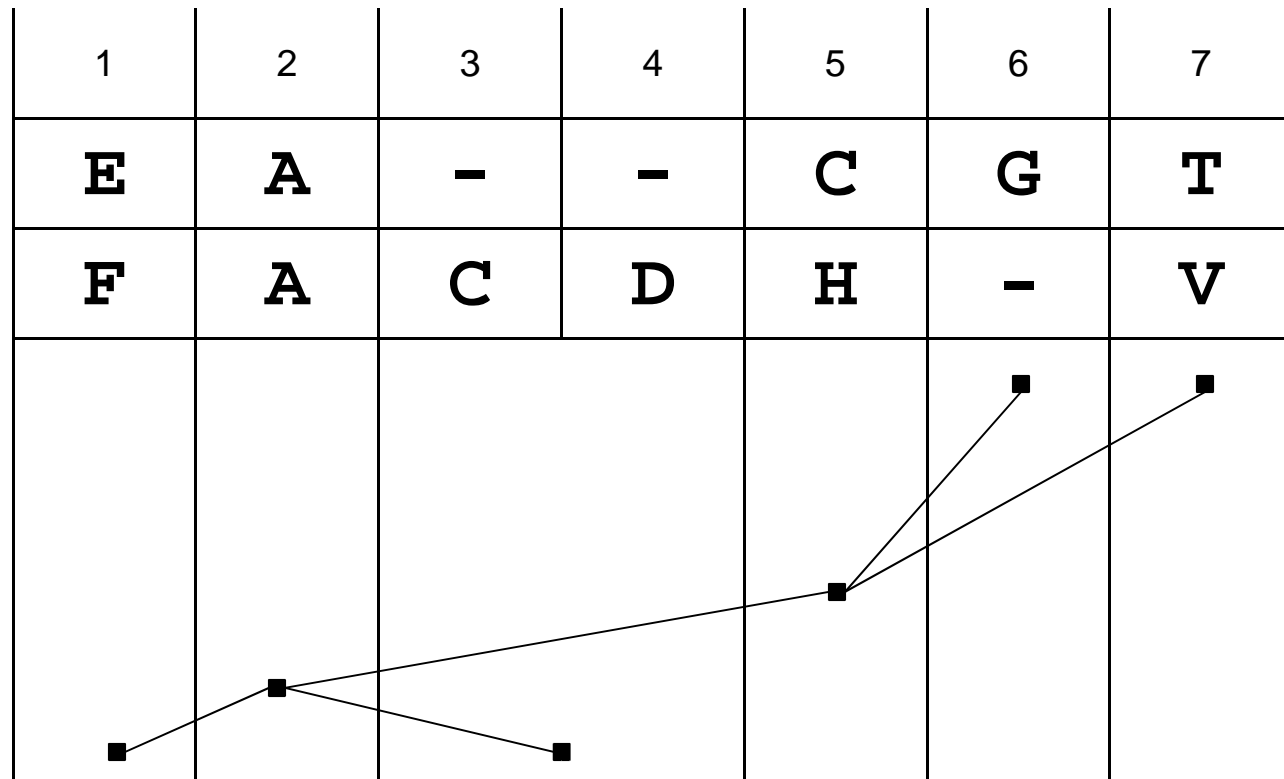
Orders Imposed

1	2	3	4	5	6	7
E	A	-	-	C	G	T
F	A	C	D	H	-	V

The following constraints are imposed by the set of operations $\{S_{E,H}(A,A), S_{A,V}(C,H), S(E,F), S(T,v), I_3, D_6\}$:

1. $S_{E,H}(A,A); S(E,F)$ due to left context **E** (pos. 2 & 1)
2. $S_{A,V}(C,H); S_{E,H}(A,A)$ due to right context of the **A**~~**A**~~ substitution (pos. 5 & 2)
3. And a few more...

Representation of the Order



Operations:

$$S_{E,H}(A,A), S_{A,V}(C,H), S(E,F), S(T,V), I_3, D_6$$

Goal

Find *alignment* and *order* which give the maximal score

Overall score is a sum of individual scores

Each position has
to be affected

Step 1: $S(\mathbf{T}, \mathbf{V})$

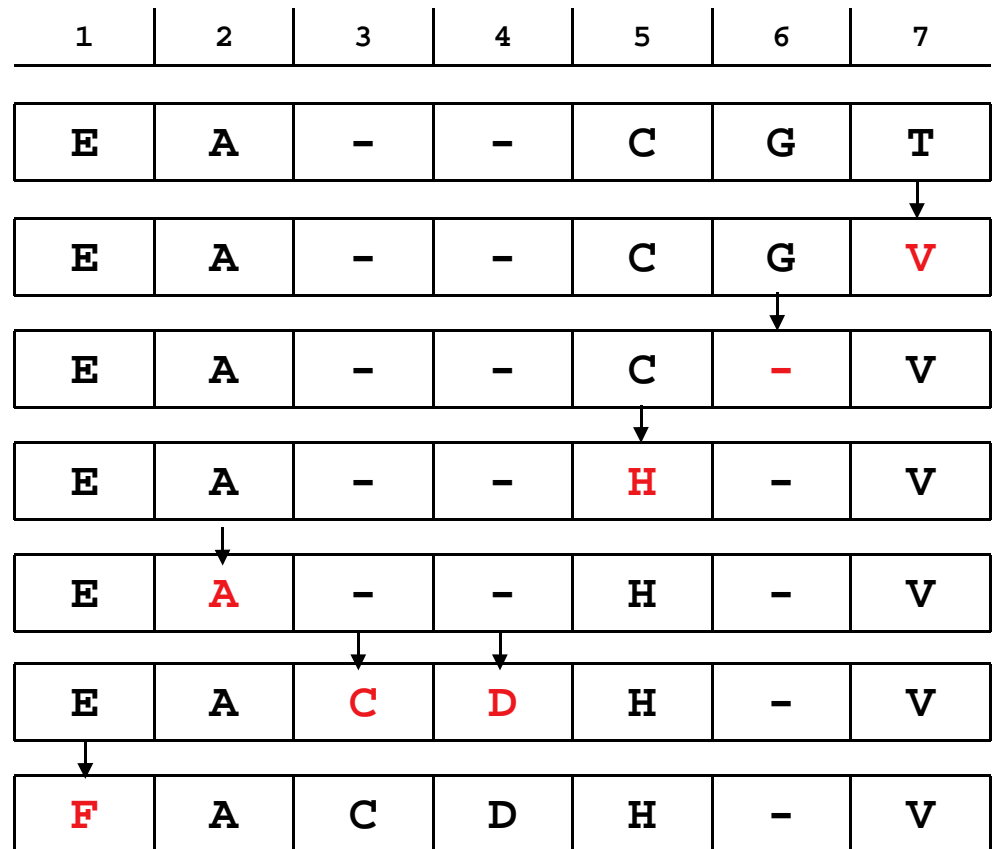
Step 2: D_6

Step 3: $S_{A,V}(\mathbf{C}, \mathbf{H})$

Step 4: $S_{E,H}(\mathbf{A}, \mathbf{A})$

Step 5: I_3

Step 6: $S(\mathbf{E}, \mathbf{F})$





Algorithms Developed

Linear time algorithm for a gap-free alignment

Quadratic time algorithm for a affine gap penalty function

Cubic time algorithm for arbitrary gap penalty

Both **local** and **global** alignment



Substitution Tables

Not enough data to create substitution tables for all possible pairs of contexts: 20^4 entries to fill in

We can group amino acids into:

- One block (i.e., *context-free*)
- Two blocks (*H, P*)
- Six blocks (biochemical properties: *basic, aromatic, aliphatic, ...*)



Experiments with COGs

Clusters of Orthologous Genes: <http://www.ncbi.nlm.nih.gov/COG>

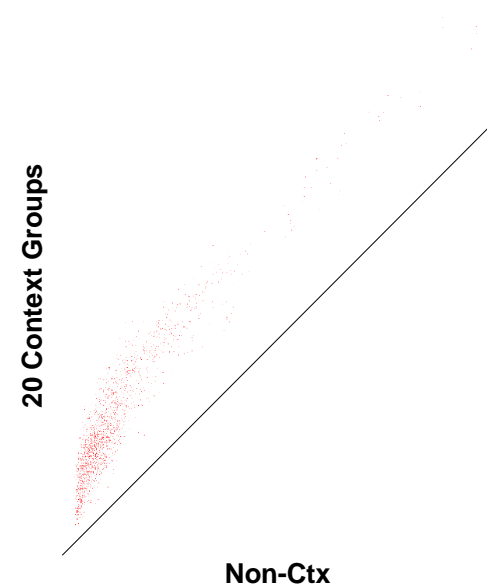
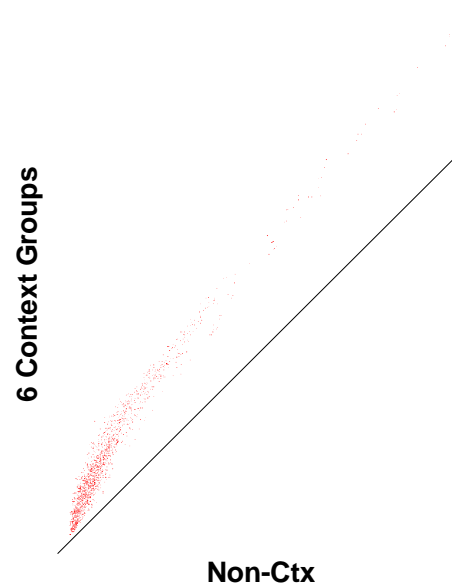
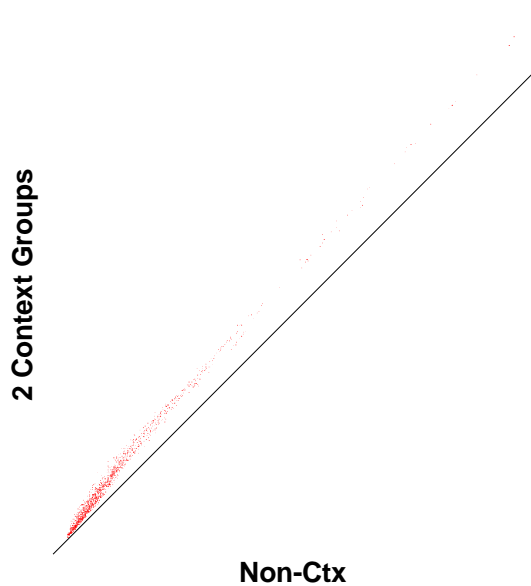
- Cluster of genes which are believed to have a common ancestor
- Created by whole-genome comparison and choosing the most similar genes

Simplified model of contextual alignment

- the score for insertion/deletion does not depend on its context
- short contexts
- Insertion has to be separated from deletion

Discrimination Power

Local alignment of COG0089 (Ribosomal proteins - large subunit)

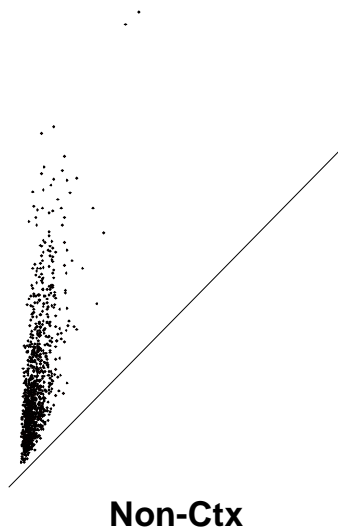


Related vs. Unrelated Proteins

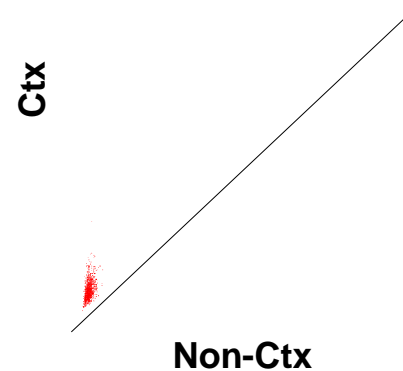
Pairs of distantly related proteins (left) have approx. 25% similarity

Unrelated proteins (right) have no statistical similarity
>1000 pairs of genes (from more than one COG)

Related proteins

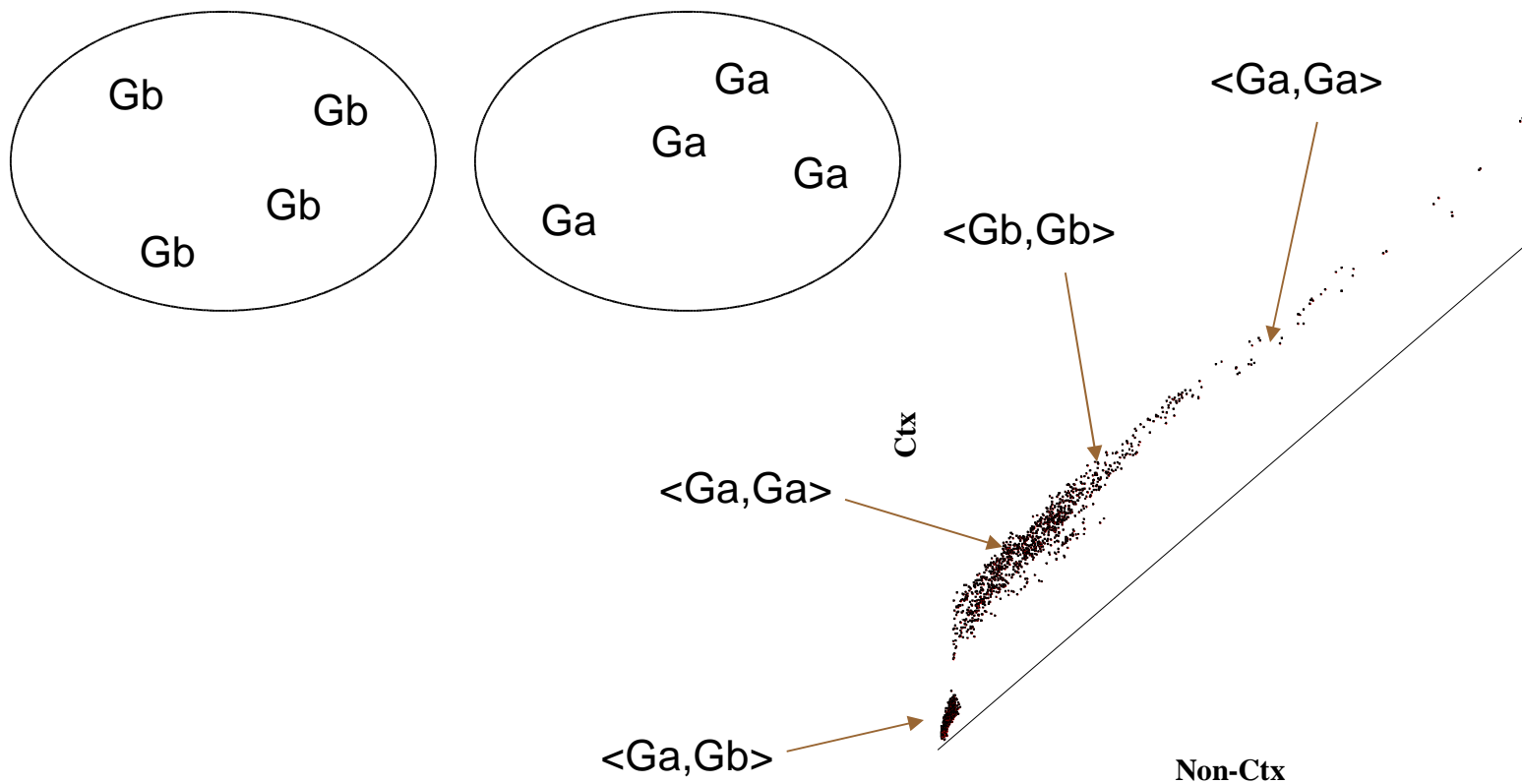


Unrelated Proteins

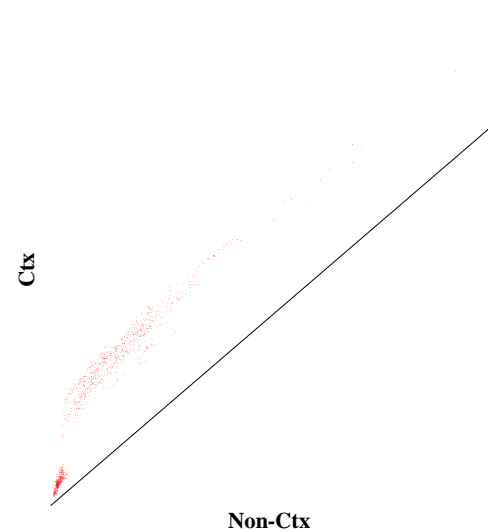
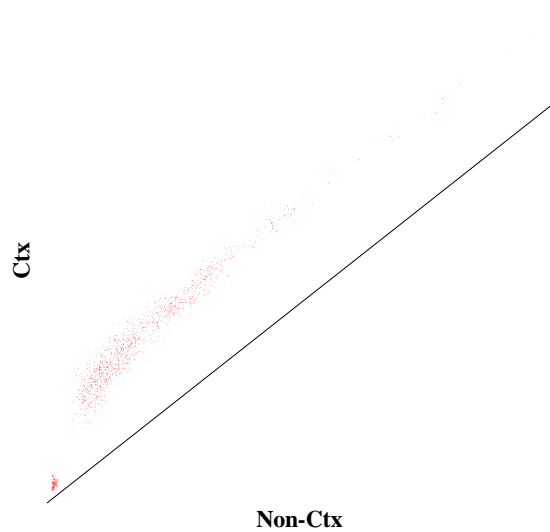
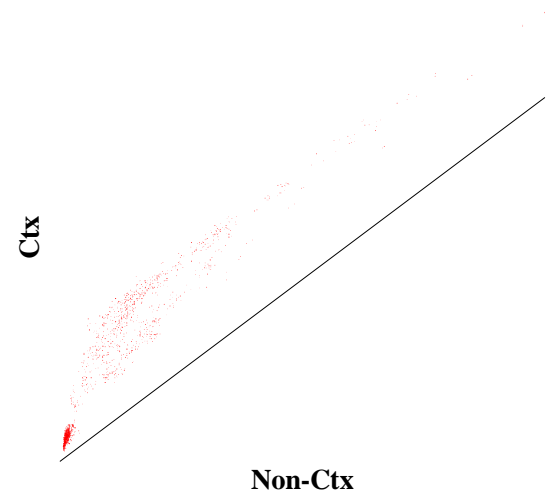
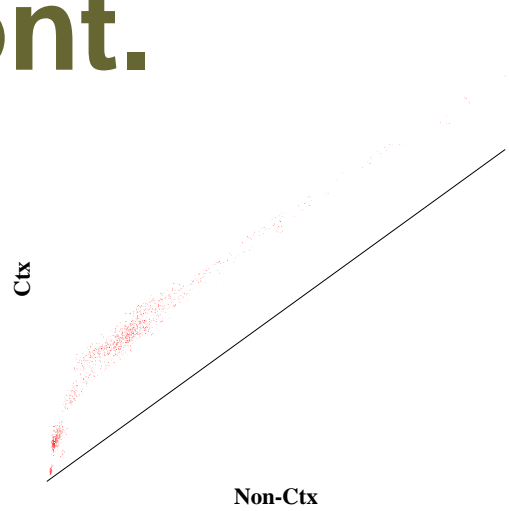


Similarity Emphasized

Two groups of closely related genes,
but with distant cross-group relation



Similarity Emphasized, cont.





Conclusions

Only close contexts were considered

The cost of insertion/deletion was context independent

Different discrimination power

- Stronger signals for similarity than non-contextual algorithm
- Detection of similarity of structure
- Grasping properties of proteins lost in non-contextual comparison



Further Applications of the Model

In phylogenetics: constructed trees are more consistent when contextual approach is used

Multiple contextual alignment: context helps in aligning orphan genes



Where to Go From Here

Context dependent indels

Longer contexts

Different kind of contexts, e.g. $i, i+1$ -
important for secondary structure of β -
sheet



Related Work

Estimation of significant context for DNA evolution in bacteriophage λ : 1 or 2 bases (*S. Tavaré and B.W. Giddings, 1989*)

Stochastic model for evolution of autocorelated DNA sequences (*A. von Haesler and M. Schöniger, 1994, 1998*)

Probabilistic model of DNA sequence evolution with context dependent rate of substitution (*L. Jensen and A.-M.K. Pedersen, 2000*)



Why Contextual?

DNA

- GC islands are highly mutable
- Transposons insert themselves in a sequence-specific manner

Proteins

- Sequence \Rightarrow structure \Rightarrow function

Algorithm

Transforms a sequence V into W

An array $T(\alpha, \beta, x)$ stores maximal score for alignment $V_1..V_\alpha$ and $W_1..W_\beta$ which ends with a substitution $V_\alpha \rightarrow W_\beta$ whose right context is x

...	$V_{\alpha-1}$	V_α
...	$W_{\beta-1}$	W_β

$$T(\alpha, \beta, x) = m \begin{cases} T(\alpha - 1, \beta - 1, V_\alpha) + s & c(S_{W_{\beta-1}, x}(V_\alpha, W_\beta)) \\ T(\alpha - 1, \beta - 1, W_\beta) + s & c(S_{V_{\alpha-1}, x}(V_\alpha, W_\beta)) \\ \dots \end{cases}$$