BIOMATH 2013
International Conference on Mathematical
Methods and Models in Biosciences
Sofia, Bulgaria, 16-21 June 2013

# Distribution at Contingency of Alignment of Two Literal Sequences

Lorentz JÄNTSCHI &
Sorana D. BOLBOACA

# OUTLINE



- INTRODUCTION & AIM
- SYMILARITY
- OUR APPROACH
- LITERAL SEQUENCE CONTINGENCY
- ALIGNMENT PROBABILITY
- PDF & CDF PLOTS
- ALIGNMENT STATISTICS
- ALIGNMENT DISTRIBUTION

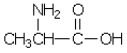
# INTRODUCTION & AIM



- Literals sequences ← encode information
  - Written language: words, phrases
  - Computer's encoding systems:
    - binary (0,1)
    - hexadecimal (0-9, A-F)
  - Genetic code:
    - DNA: A, G, C, T
    - RNA: A, G, C, U
  - Proteins: amino acids

| | | | | | |
|---|---|---|---|---|---|
| Octal | 3 | 4 | 6 | 4 | 2 |
| Binary | 011 | 100 | 110 | 100 | 010 |
| Binary | 0011 | 1001 | 1010 | 0010 | |
| Hexadecimal | 3 | 9 | A | 2 | |

sorana.academicdirect.ro/pages/collagen/amino_acids/aa_web.htm

| Amino Acid | Systemic name | Symbol 3 | Symbol 1 | Structure | Molecular Formula | Molar weight |
|---|---|---|---|---|---|---|
| Alanine | (S)-2-Aminopropanoic Acid | Ala | A |  | $C_3H_7NO_2$ | 89.09 |
| Arginine | 2-Amino-5-(Diaminomethylidene Amino) Pentanoic Acid | Arg | R |  | $C_6H_{14}N_4O_2$ | 174.20 |

# Aim



- Hypothesis: distribution of alignments could provide useful information about the chance that a certain alignment occur or not by chance.
- Literals sequences alignment: analysis of similarity
  - Similarity of two texts (e.g. copy-paste issue)
  - Similarity of two computer encodings (e.g. copy rights)
  - Similarity of two genetic codes (biological compatibility; phylogeny)
  - Similarity of two proteins (biological functionality replacements)

# TYPES OF SYMILARITIES



- Exact matching
  - Gives probabilities of non-driven (or by chance) match (categorical data [1])
  - Statistical issues addressed
- Using insertions, deletions and shifts
  - Gives minimal set of insertions, deletions and shifts to match two (or more) sequences → costs (energetic for instance) for matching
  - Meta-heuristic issues addressed

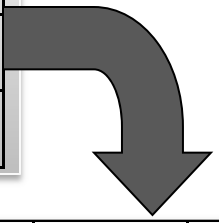


1. Wongsuphasawat K, Plaisant C, Taieb-Maimon M, Shneiderman B. Querying event sequences by exact match or similarity search: Design and empirical evaluation. Interacting with Computers 2012;24(2):55-68.

# Our Approach: Exact Match

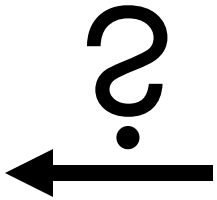


| j | 1 | 2 | 3 | 4 | … | n-4 | n-3 | n-2 | n-1 | n |
|---|---|---|---|---|---|---|---|---|---|---|
| Seq1 | A | C | C | G | … | T | A | G | A | C |
| Seq2 | C | A | C | T | … | A | A | G | C | A |
| Seq1Seq2 | AC | CA | CC | GT | … | TA | AA | GG | AC | CA |
| ? | ✗ | ✗ | ✓ | ✗ | … | ✗ | ✓ | ✓ | ✗ | ✗ |



| R\C | A | C | G | T | Σ |
|---|---|---|---|---|---|
| A | **ΣAA** | ΣAC | ΣAG | ΣAT | |
| C | ΣCA | **ΣCC** | ΣCG | ΣCT | |
| G | ΣGA | ΣGC | **ΣGG** | ΣGT | |
| T | ΣTA | ΣTC | ΣTG | **ΣTT** | |
| Σ | | | | | |

Perfect alignment:
ΣAA, ΣCC, ΣGG, ΣTT
(main diagonal)




Alignment ratio:
(ΣAA+ΣCC+ΣGG+ΣTT)/n

# Literals Sequences Contingency (General Case)



| R\C | $c_1$ | … | $c_q$ | Σ |
|---|---|---|---|---|
| $r_1$ | $a_{11}$ | … | $a_{1q}$ | |
| … | … | … | … | |
| $r_q$ | $a_{q1}$ | … | $a_{qq}$ | |
| Σ | | | | |

For two sequences of q literals with size n, the alignment is defined by

$$SqA(n,q) \overset{def}{=} \sum_{k=1}^{q} a_{k,k}$$

SqA(n,q) is a distribution of integers ranging from 0 (no alignment) to n (perfect alignment):

$$SqA_F(0;n,q) \quad ... \quad SqA_F(i;n,q) \quad ... \quad SqA_F(n;n,q)$$

$$\{0, \quad ..., \quad i, \quad ..., \quad n\}$$

# Probability of an Certain Alignment



- Imposed conditions (alignment of $U_i$ vs $V_i$, $1 \le i \le n$):
  - $\Sigma_j a_{j,k} \neq 0$ for $1 \le k \le q$ (all letters of the alphabet (or contingency) are present at least once in the second sequence of literals
  - $\Sigma_k a_{j,k} \neq 0$ for $1 \le j \le q$ (all letters of the alphabet (or contingency) are present at least once in the first sequence of literals
- A math calculation gives their frequencies:

$$SqA_F(i;n,q) = \begin{pmatrix} n-i+q^2-q-1 \\ q^2-q-1 \end{pmatrix} \begin{pmatrix} i+q-1 \\ q-1 \end{pmatrix}$$

# CDF & PDF



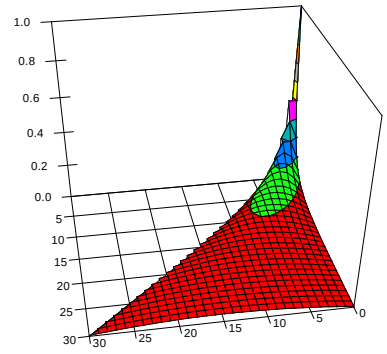
$$\mathrm{SqA}_{\mathrm{PDF}}(i;n,q)=\frac{\Gamma(q^2)}{\Gamma(q^2-q)\Gamma(q)}\frac{\Gamma(n+1)}{\Gamma(n+q^2)}\cdot\frac{\Gamma(i+q)}{\Gamma(i+1)}\frac{\Gamma(n-i+q^2-q)}{\Gamma(n-i+1)}$$

$$\mathrm{SqA}_{\mathrm{CDF}}(i;n,q)=\frac{\Gamma(q^2)}{\Gamma(q^2-q)\Gamma(q)}\frac{\Gamma(n+1)}{\Gamma(n+q^2)}\sum_{j=0}^{i}\frac{\Gamma(j+q)}{\Gamma(j+1)}\frac{\Gamma(n-j+q^2-q)}{\Gamma(n-j+1)}$$

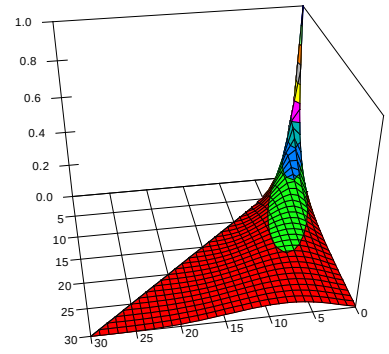
CDF: no close form!

# PDF & CDF PLOTS (DENSITY OF PROBABILITY VS. 'ALIGNED' (Σ MAIN DIAGONAL) AND 'TO BE ALIGNED' (N)

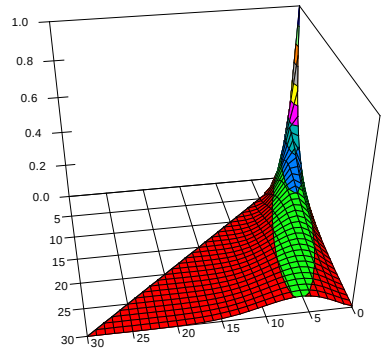




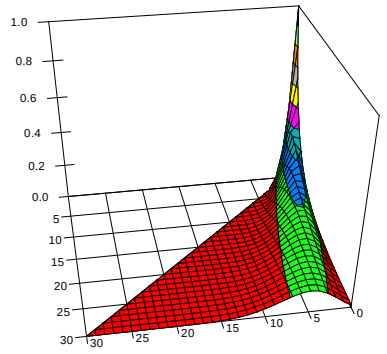

PDF2 (q = 2)



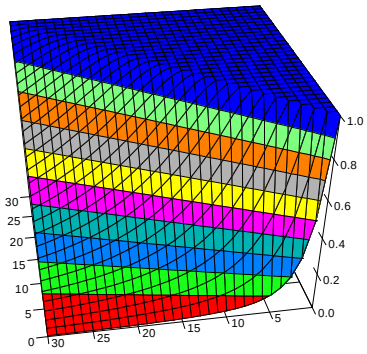

PDF3 (q = 3)



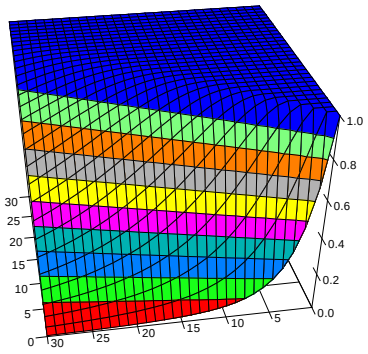

PDF4 (q = 4)



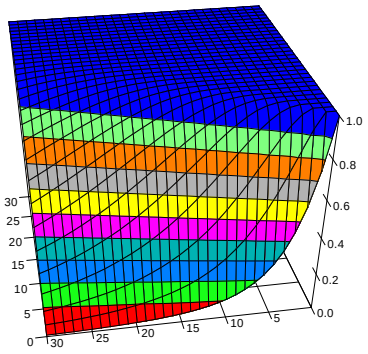

PDF5 (q = 5)



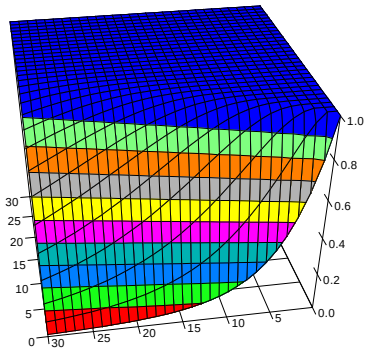

CDF2 (q = 2)




CDF3 (q = 3)




CDF4 (q = 4)




CDF5 (q = 5)

# PERFECT ALIGNMENT STATISTICS



Mean

$$\mu(n,q) = \frac{n}{q}$$

Standard deviation

$$\sigma^2(n,q) = n(n+q^2)\frac{(q-1)}{q^2(q^2+1)}$$

Mode

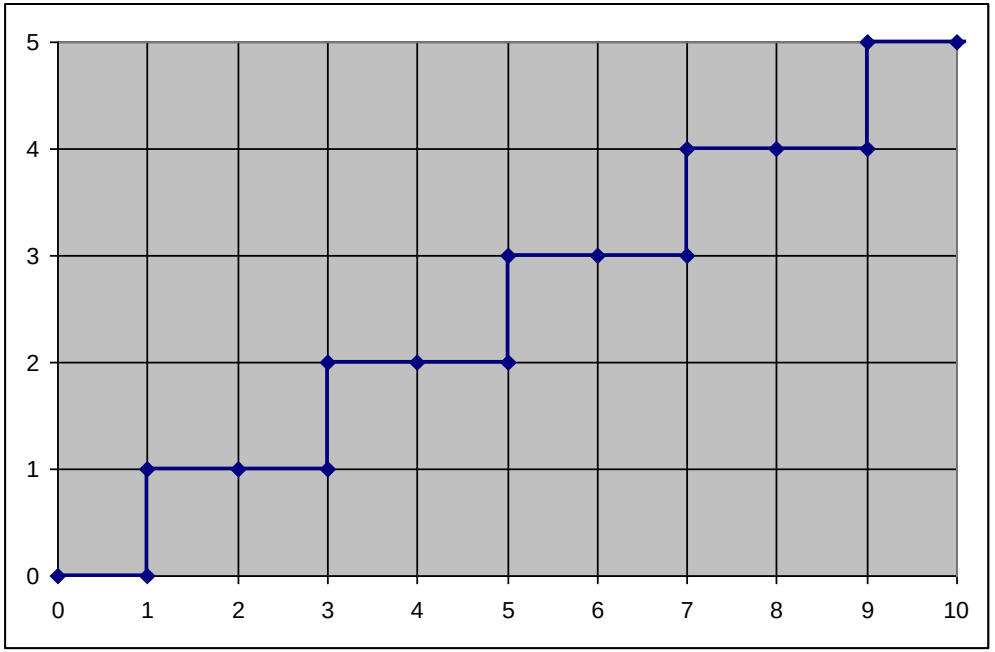
for q=2: When n is odd are two modes (figure above)

for q=3: when $7 \cdot i = 2 \cdot n - 5$ (n = 20, 27, 41, 55, …)

for q=4: when $14 \cdot i = 3 \cdot n - 11$ (n = 27, 83, …)

for q=5: when $23 \cdot i = 4 \cdot n - 19$ (n = 91, …)

in general: when $(q^2-2) \cdot i = (q-1) \cdot n - (q^2-q-1)$




q = 2 (encoding with 0 and 1)

Mode: alignment with highest probability to be observed by chance

Figure: 'aligned' (by chance) vs. 'to be aligned' (sequences size)

# Other Distribution Statistics



$$\sigma^2(n,q) = n(n+q^2)\frac{(q-1)}{q^2(q^2+1)}$$

$$\mu_3(n,q) = n(n+q^2)(2n+q^2)\frac{(q-1)(q-2)}{q^3(q^2+1)(q^2+2)}$$

$$\gamma_1(n,q) = 2\frac{n+q^2/2}{\sqrt{n(n+q^2)}}\frac{q-2}{\sqrt{q-1}}\frac{\sqrt{q^2+1}}{q^2+2}$$

$$\mu_4(n,q) = \frac{n(n+q^2)(q-1)\left(q^6+3(n-2)q^5+(3n+5)q^4+3n(n-6)q^3+3n(n+6)q^2-18n^2(q-1)\right)}{q^4(q^2+1)(q^2+2)(q^2+3)}$$

$$\beta_2(n,q) = \frac{(q^2+1)\left(q^6+3(n-2)q^5+(3n+5)q^4+3n(n-6)q^3+3n(n+6)q^2-18n^2(q-1)\right)}{n(n+q^2)(q^2+2)(q^2+3)(q-1)}$$

$$\gamma_2(n,q) = \beta_2(n,q) - 3$$

# ALIGNMENT DISTRIBUTION APPLICATIONS



**"Alignment distribution":**

- If Alignment ratio greater than the threshold of 95% (InvCDF95) then with a risk smaller than 5% being in error the alignment of the sequences is not by chance.

# Alignment Distribution Applications



**Important points for q=4 (gene sequence alignment):**

- Thresholds to reject the alignment by chance

| n | Alignment % |
|---|---|
| 8 | 70 |
| 13 | 60 |
| 21 | 55 |
| 39 | 50 |
| 282 | 45 |
| $\rightarrow\infty$ | 44 |

# Thank your for your attention