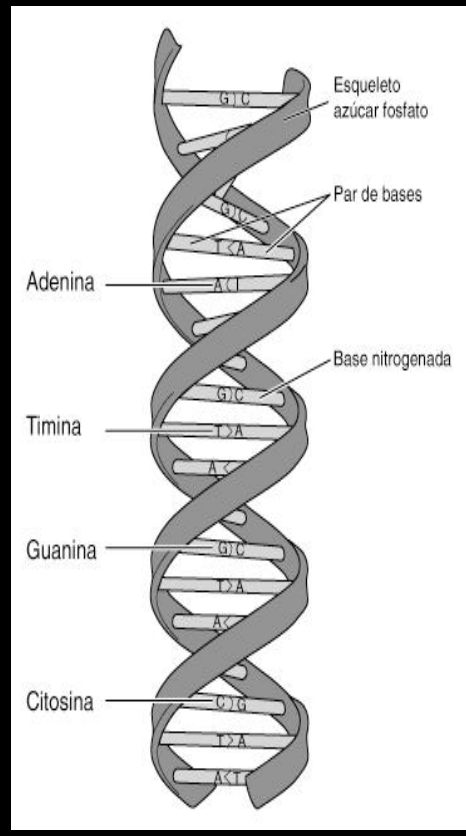
Compresión de la secuencia del ADN

Cristina Galindo Perrino
3º Telecomunicaciones
U.S.P CEU

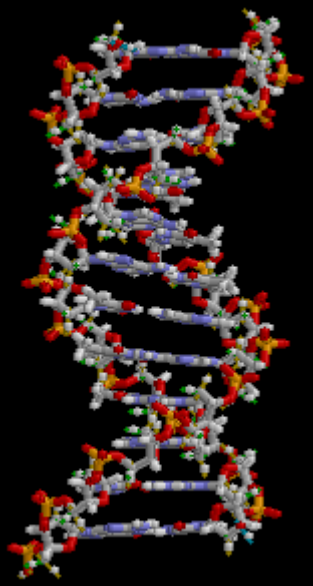
- IEEE SIGNAL PROCESSING MAGAZINE
- Enero 2007
- Gergely Korodi, Ioan Tabus, Jorma Rissanen y Jaakko Astola

ADN

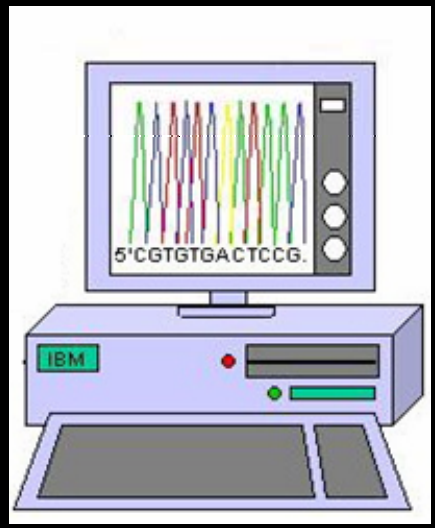
- Ácido desoxirribonucleico.
- El ADN es un polímero formado por cuatro tipos de nucleótidos, diferenciados por sus bases nitrogenadas divididas en dos grupos: adenina (**A**), guanina (**G**) y citosina (**C**), timina (**T**).
- Un organismo tiene en cada una de sus células la misma secuencia de ADN, estructurando así el genoma del individuo.



Secuencia



- Genes (forman parte de los exones)
- $4^3 = 64$ tripletas de nucleótidos
- S = {A, C, G, T}
- S = {0, 1, 2, 3}



Compresión

- En organismos con un tamaño grande de ADN, aparece la redundancia.
- Dos tipos:
  - Redundancia Directa

  Ej: ACTTGTC...ACTTA…
  - Palíndromos complementarios

  $x_i,\ldots,x_{i+n}$ $y_p,\ldots,y_{p+n}$

  Ej: ACTTGTC

  GACAAGT
- Los métodos de compresión, expanden.

Repeticiones y cambios

- Se propagan de generación en generación.
- Repeticiones de patrones: aproximada y exacta.
  - ACTTGTC siguiente generación ACTCGTC
  - ACTTGTC siguiente generación ACTTGTC

Repeticiones y cambios

- Cambios: sustitución y eliminación o inserción de símbolos
  - ACTTGTC → ACTXGTC
  - ACTTGTC → ACTTGATC
- Estos cambios afectan en la compresión.
- Igualdad o repetición significativa biológicamente.
  - ACT —... 30 bases...→ ACT
  - ACT —...10000 bases...→ ACT

Evolución de los métodos de compresión

- 1960-1990
- Zip, basado en el algoritmo Lempel-Ziv y el bzip2.
- Biocompress y Biocompress2, basados en el LZSS (dentro de la familia Lempel-Ziv).
- C-fact
- GenCompress I
- DNA Compress, usa el PatternHunter.
- NMLComp. Distancia Hamming
  - d_H(ACT<u>T</u>GTC, ACT<u>C</u>GTC) = 1

Modelo NML

- Divido la secuencia en segmentos (de n en n).

$$\left\{ \underline{y}(k) = \underline{y}_{kn+1}^{kn+n} \stackrel{\text{def}}{=} y_{kn+1} \cdots y_{kn+n} \mid k = 0, \ldots, \left\lfloor \frac{N}{n} \right\rfloor - 1 \right\}$$

Ej: ACTTGTCTACGGAATCTGCT…

5 5 5

Modelo NML

- Construcción del diccionario.

$$W = \bigcup_{p=0}^{kn-n} \left\{ \underline{x}_{p+1}^{p+n}, R\left(\underline{x}_{p+1}^{p+n}\right) \right\}$$

Ej: ACTTGTCTACGGAATCTGCT…

W = { ACTTG, CAAGT
TCTAC, GTAGA
… … }

Modelo NML

- Busco en el diccionario la palabra de distancia mínima (Hamming) por la que sustituyo la $\underline{y}$.

$$\underline{x} = \arg \min_{\underline{z} \in W} ||\underline{z} - \underline{y}||_H$$

Ej: si $\underline{y}$ = TCTGC

d_H = (TCTGC, ACTTG) = 3

d_H = (TCTGC, CAAGT) = 4

d_H = (TCTGC, TCTAC) = 1

$\underline{x}$ = TCTAC

Modelo NML

- Calculo probabilidad de conociendo la palabra del diccionario $\underline{x}$, sepa la palabra que quería sustituir $\underline{y}$.

$$P(\underline{y}|\underline{x}) = \frac{(|S|-1)^{m-n}\left(\frac{m}{n}\right)^{m}\left(\frac{n-m}{n}\right)^{n-m}}{\sum_{j=T(n)}^{n}\binom{n}{j}\left(\frac{j}{n}\right)^{j}\left(\frac{n-j}{n}\right)^{n-j}}$$

Modelo NML

- Busco las subsecuencias de longitud r que coincida.

$$M_{i,r}(W) = \left\{ \underline{x}_1^n \in W \mid \underline{x}_{i+1}^{i+r} = \underline{y}_{i+1}^{i+r} \right\}$$

0 1 2 3 4

Ej: TCTGC

3

$$W = \left\{ \begin{array}{ll} ACTTG, & CAAGT \\ TCTAC, & GTAGA \\ \ldots & \ldots \end{array} \right\}$$

$M_{0,3} = \{1\}$

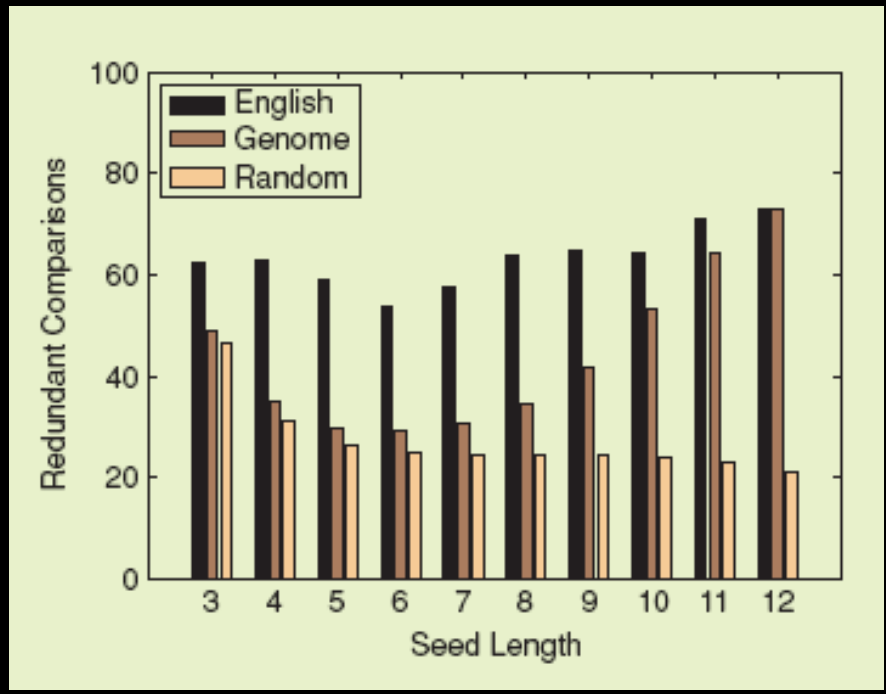
Modelo NML

- Minimizar la distancia Hamming, esta vez a partir del conjunto de las subsecuencias anteriores.

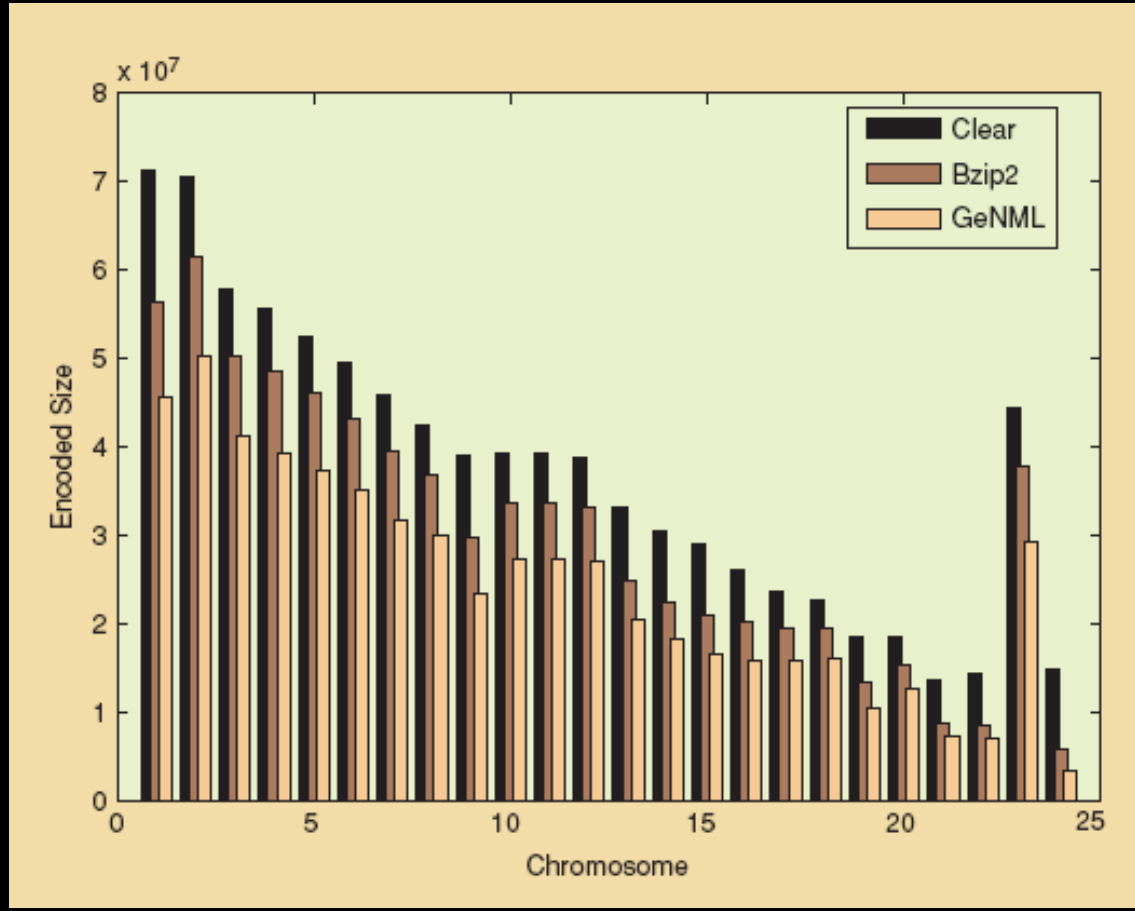
$$D_r = \bigcup_{i=0}^{n-r} M_{i,r}(W)$$

$$\underline{x}' = \arg\min_{\underline{z} \in D_r} |\underline{z} - \underline{y}||_H$$

Proporción de la redundancia en: la secuencia del ADN humano, un texto en inglés y una secuencia aleatoria



Comparación de los modelos de compresión



Conclusión

- Menor espacio en disco y en transmisión.
- Posibilidad de comparar los genomas completos.
- Significado biológico extraído de los datos de la redundancia.