Automated technology for revealing of latent periodicities in DNA sequences

Chaley Mariya, Kutyrkin Vladimir, Tulbasheva Gayane,
Teplukhina Elena, Nazipova Nafisa

Institute of Mathematical Problems of Biology RAS

Chaley M.; Kutyrkin V.; Tulbasheva G.; Teplukhina E. Nazipova N. HeteroGenome: database of genome periodicity. *Database.* 2014, Vol. 2014, P. 1-18.

Definitions

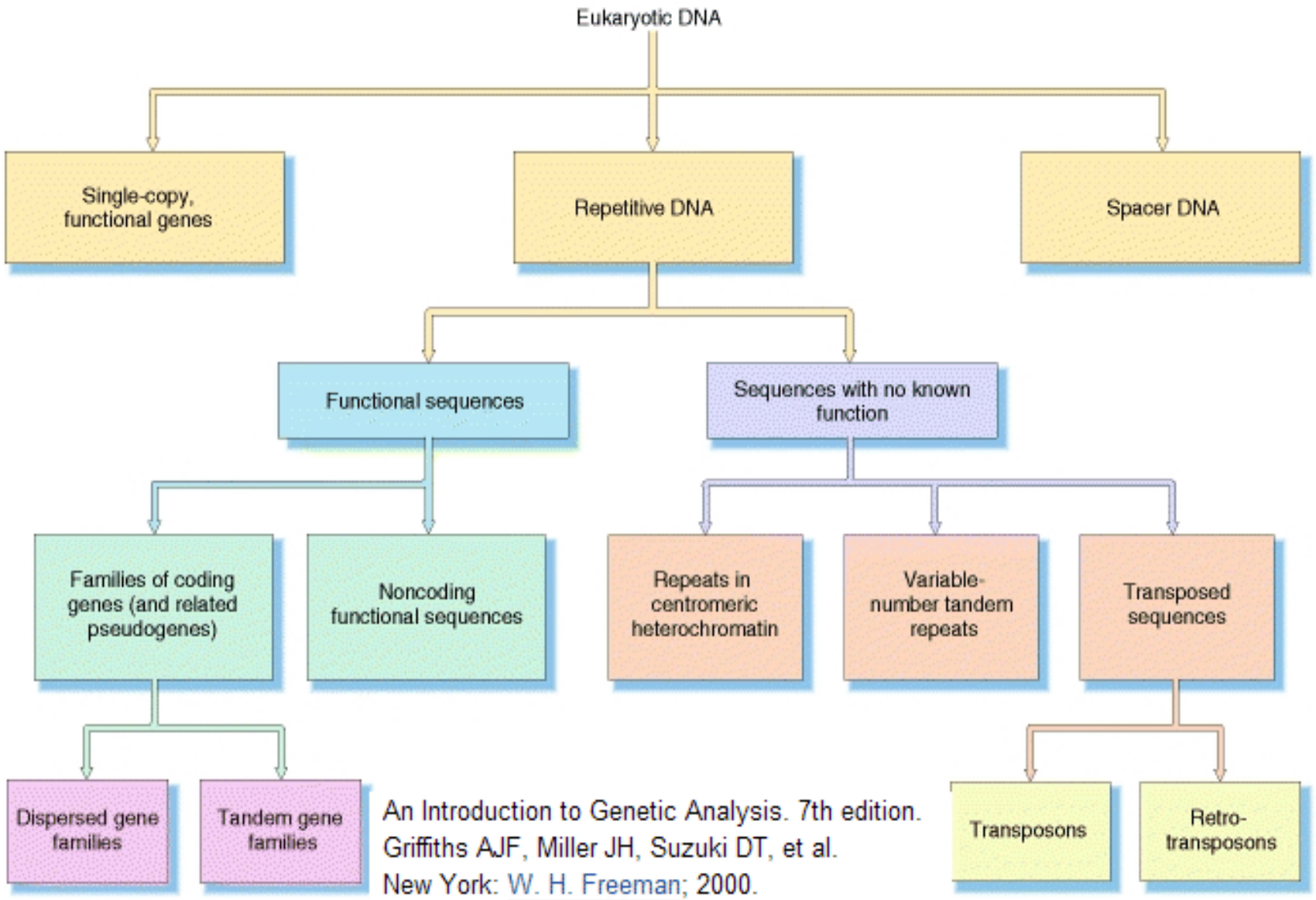
Latent periodicity (pattern length L=7, copy number R=7)

…tacc ATCGCT°ATGGCT°ATCGCT°ATCCCT°ATCGGCT°ATCCT ATCGCT tgcaa…

Structural classification of repetitive DNA and DNA sequence variation types:

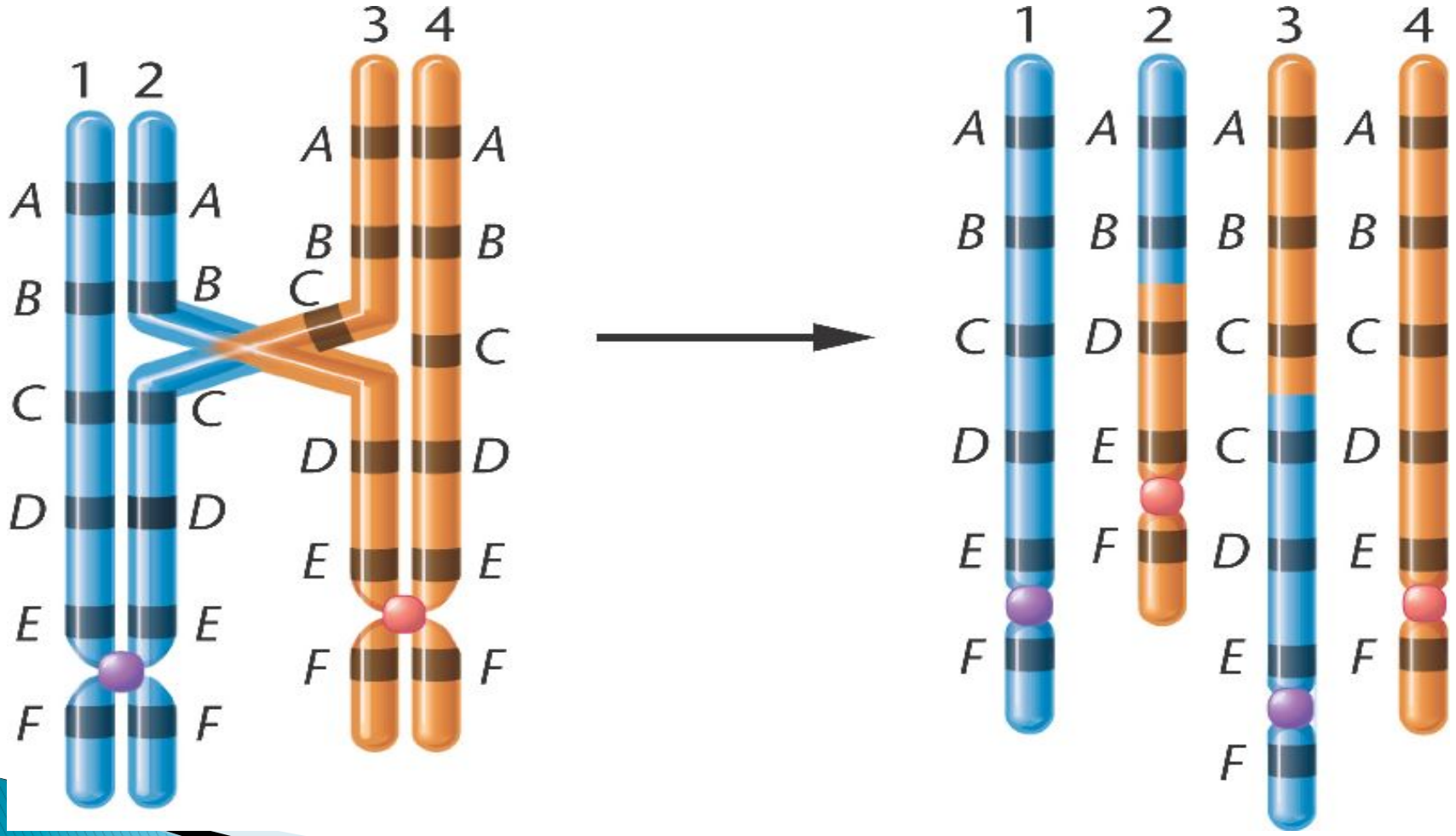
- Microsatellites (SSRs, Simple Sequence Repeats) ($1 < L \leq 10$)
- Minisatellites (VNTRs, Variable Number of Tandem Repeats) ($10 < L \leq 100$)
- Macrosatellites (CNVs, Copy Number Variations) ($100 < L \leq 400$)
- Megasatellites (CNVs, Copy Number Variations) ($400 < L$)

Functional Classification of Eukaryotic DNA

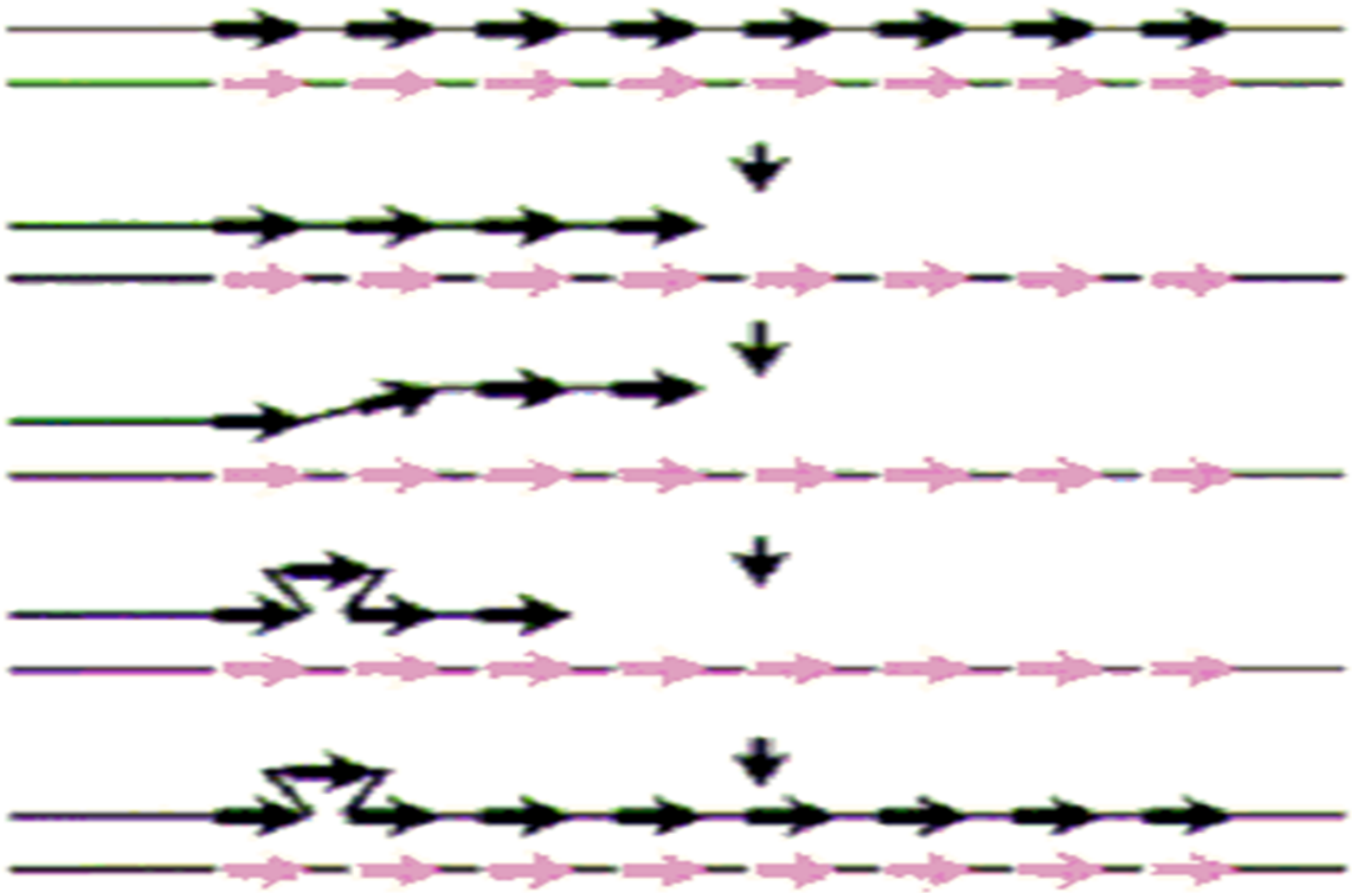


An Introduction to Genetic Analysis. 7th edition.
Griffiths AJF, Miller JH, Suzuki DT, et al.
New York: W. H. Freeman; 2000.

Unequal crossing over is a duplication event which cause emergence of repeats in genome



RNA polymerase slippage is one of the mechanism of pattern multiplication



Localization of repetitive DNA types in the chromosome (example of sugar beet)

Intercalary tandem repeats

Centromere associated tandem repeat

Telomeric and sub-Telomeric repeats

Dispersed tandem repeats

Dispersed Ty-1-copia-like retroelements LTR and microsatellites

Single and low-copy sequences Including genes

Our method basics

…tacc ATCGCT ATGGCT ATCGCT ATCCCT ATCGGCT ATCCT ATCGCT ATTG tgcaa…

$$\boldsymbol{\pi} = (\pi_j^i)_L^4, alp = \{\mathrm{A},\mathrm{T},\mathrm{G},\mathrm{C}\}$$

$$\boldsymbol{\pi} = \begin{pmatrix} 7 & 1 & 0 & 0 & 0 & 0 \\ 1 & 7 & 2 & 0 & 0 & 6 \\ 0 & 0 & 1 & 6 & 1 & 0 \\ 0 & 0 & 5 & 2 & 6 & 1 \end{pmatrix}$$

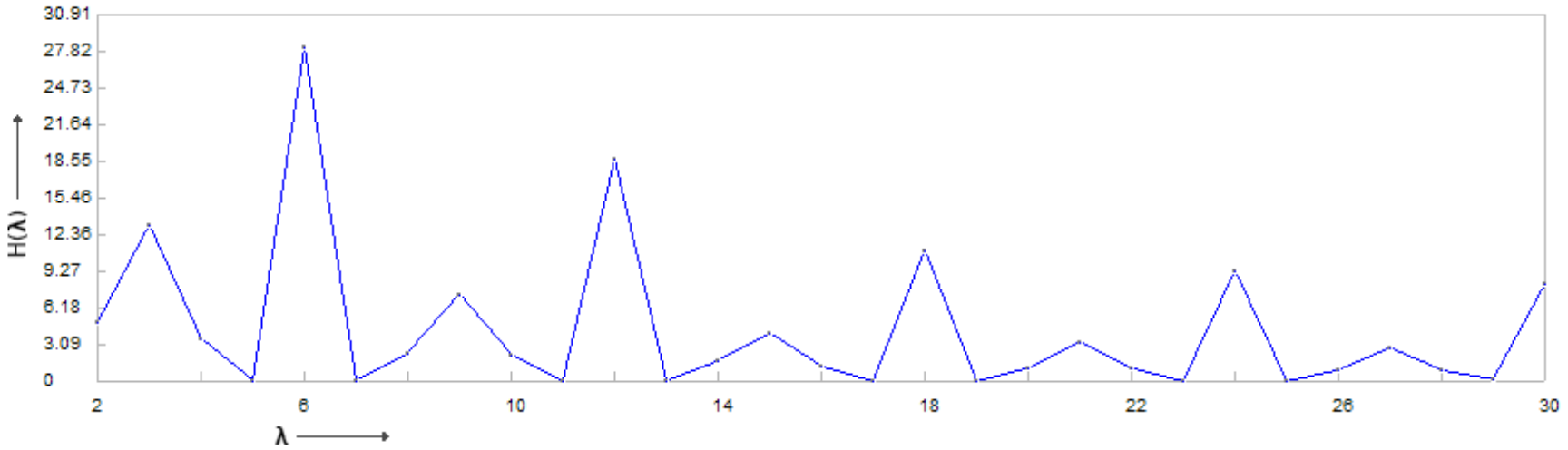
$$pl(L) = \frac{1}{L}\sum_{j=1}^{L} \max(\pi_j^i : i = 1,2,3,4)$$

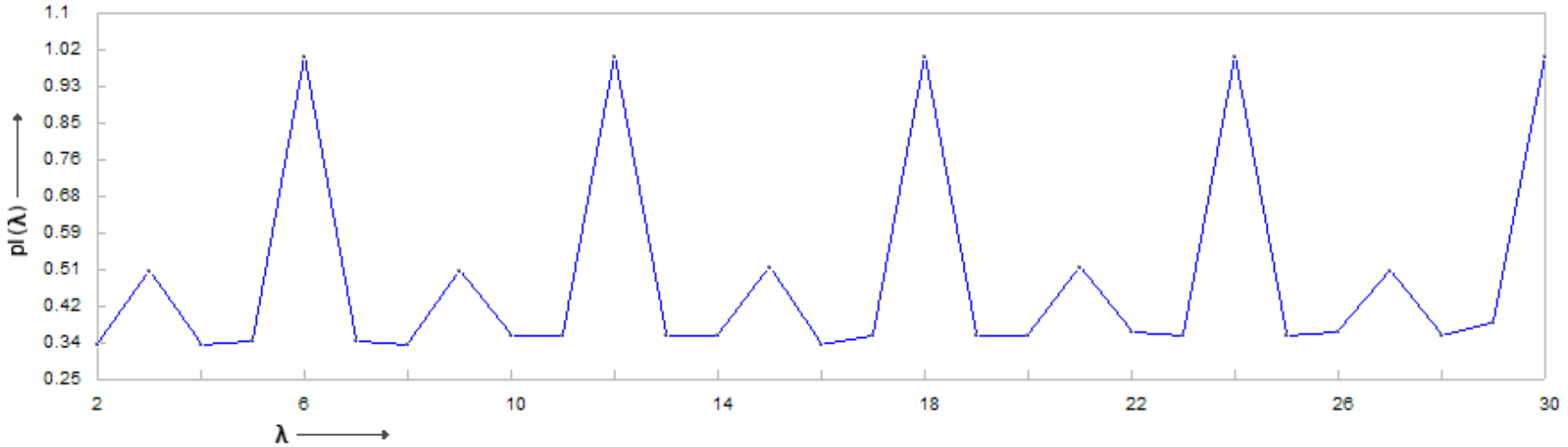
$$HL(L) = \frac{N}{L\chi^2_{crit}(\alpha, 3(L-1))} \sum_{j=1}^{L}\sum_{i=1}^{4} \frac{(\pi_j^i - p^i)^2}{p^i(1-p^i)}, \quad \alpha = 10^{-6}$$

$$E(L) = -\sum_{j=1}^{L} p^i \log_4 p^i, \qquad p^i = \frac{1}{L}\sum_{j=1}^{L} \pi_j^i, i = 1,2,3,4.$$

Joint use of the two criteria allows determining of the true length of pattern

1 - 432, λ=6, PL=1, HL=28.1, RL=72 | 2 <= λ <= 30 | Define new range | Help

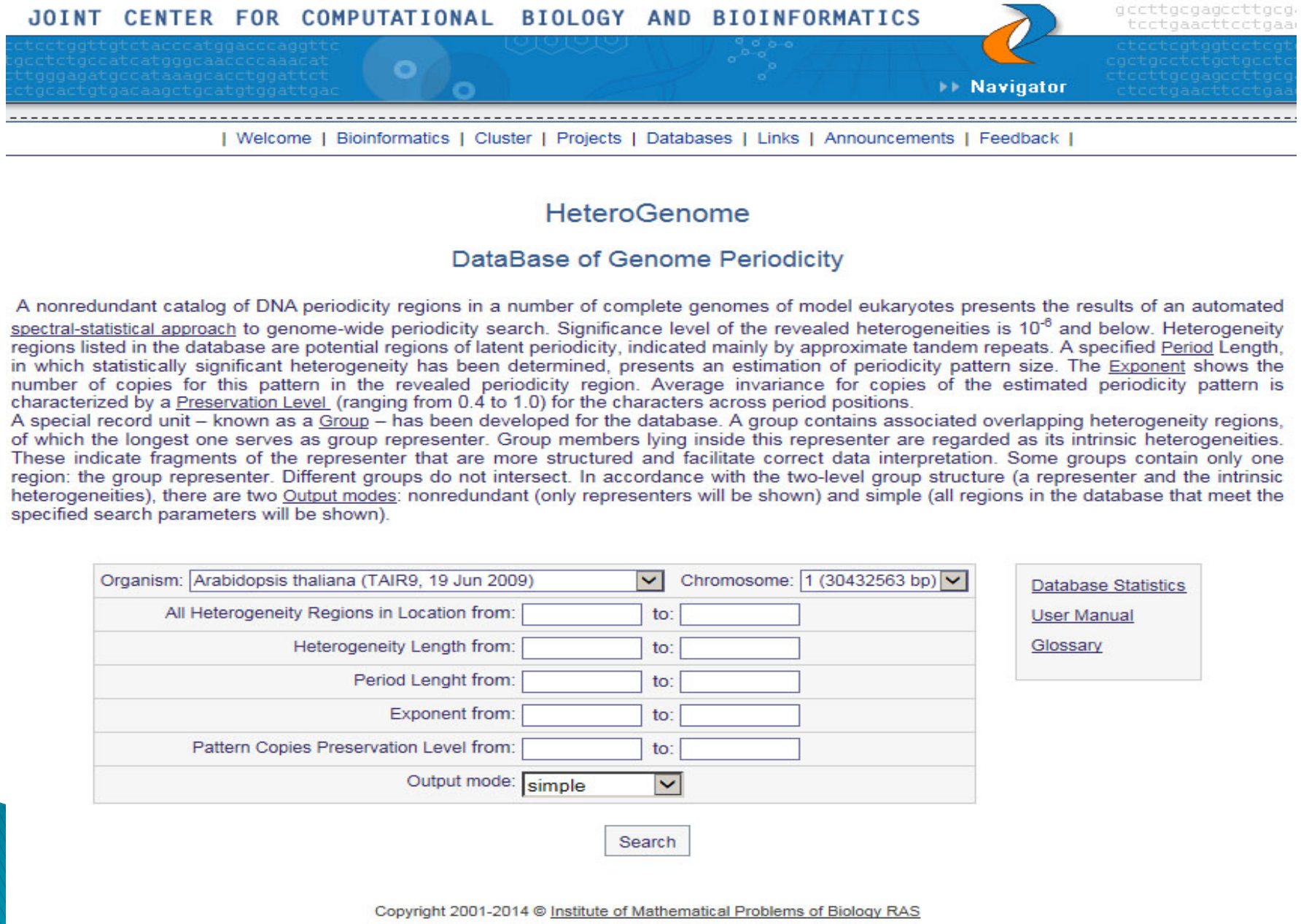




Main blocks of automated technology

- Genome scanning with all the values of L ($2 \leq L \leq 4000$).
- Analysis of the set of identified regions: elimination of full entrances, additional analysis of intersections and fragments with multiple values of L lying in the close proximity to each other. Staged reduction of redundancy.
- Construction of two-level structures – the representatives of groups and the groups of concerted internal repeats.

http://www.jcbi.ru/lp_baze/



JOINT CENTER FOR COMPUTATIONAL BIOLOGY AND BIOINFORMATICS

Navigator

| Welcome | Bioinformatics | Cluster | Projects | Databases | Links | Announcements | Feedback |

HeteroGenome

Database of Genome Periodicity

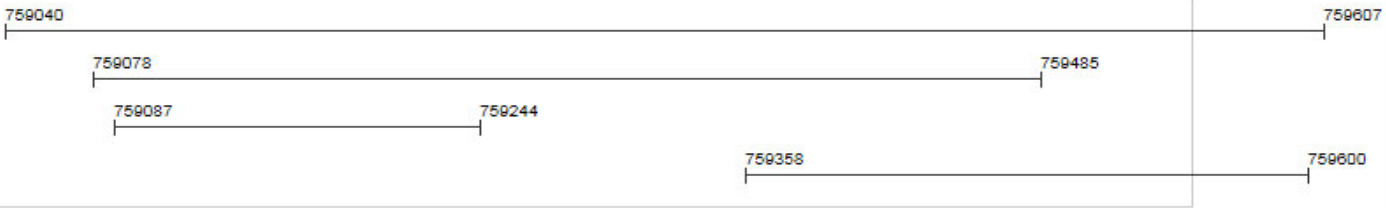
Organism: Caenorhabditis elegans (WS150, 21 Oct 2005) Chromosome: I (15072418 bp) Statistics Home

Location	Length	Period	Exponent	Preservation Level	H-spectrum value
759040 - 759607	568	35	16.23	0.61	2.23

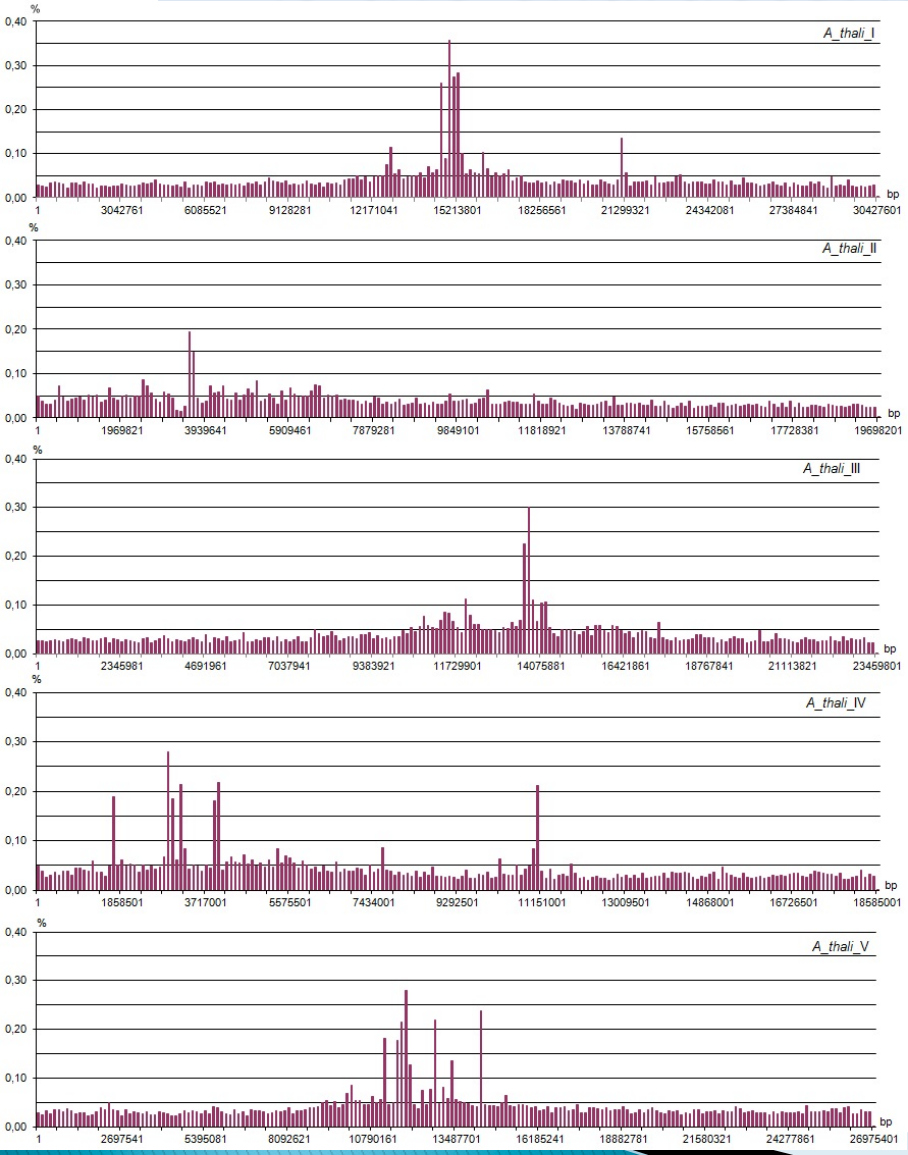
Show spectra Show sequence Sequence Viewer

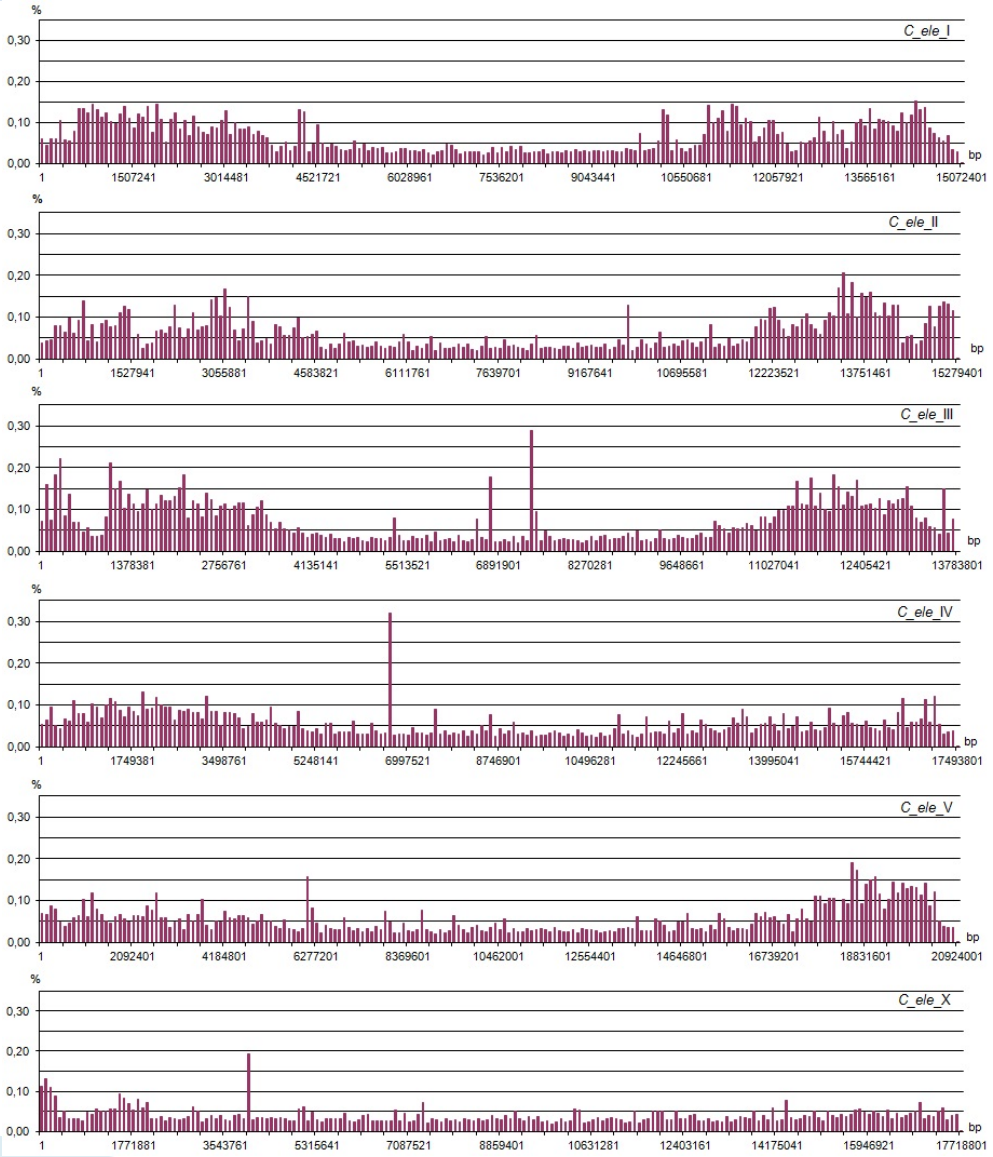
INTRINSIC HETEROGENEITIES:

Location	Length	Period	Exponent	Preservation Level	H-spectrum value			
759078 - 759485	408	34	12	0.62	1.94	Show spectra	Show sequence	Sequence Viewer
759087 - 759244	158	34	4.65	0.96	2.45	Show spectra	Show sequence	Sequence Viewer
759358 - 759600	243	106	2.29	0.92	1.57	Show spectra	Show sequence	Sequence Viewer



Distribution of repeats along a chromosome is the characteristic of the chromosome





	Length, bp	Share of Genome Length, %			
		micro	mini	mega	all
A. thaliana	**119146348**	**1.86**	**4.46**	**2.02**	**8.34**
Chr I	30427671	1.86	4.25	2.47	8.58
Chr II	19698289	1.88	4.83	1.25	7.96
Chr III	23459830	1.83	4.37	1.76	7.96
Chr IV	18585056	1.86	4.59	2.29	8.74
Chr V	26975502	1.88	4.26	2.31	8.45
C. elegans	**100272607**	**2.04**	**7.26**	**2.50**	**11.80**
Chr I	15072434	2.25	8.13	3.04	13.42
Chr II	15279421	2.03	7.66	2.93	12.62
Chr III	13783801	2.07	8.93	3.57	14.57
Chr IV	17493829	2.08	6.92	2.18	11.18
Chr V	20924180	1.93	6.90	2.36	11.18
Chr X	17718942	1.90	5.02	0.91	7.83
D. melanogaster	**133880608**	**2.93**	**4.19**	**1.34**	**8.45**
Chr 2L	23513712	2.70	3.68	0.97	7.35
Chr 2R	25286936	2.75	3.80	1.47	8.02
Chr 3L	28110227	2.79	3.95	1.13	7.86
Chr 3R	32079331	2.92	4.03	1.28	8.23
Chr 4	1348131	2.32	4.30	1.45	8.07
Chr X	23542271	4.09	5.37	1.73	11.18

Database statistics for 3 higher organisms showing the relative uniformity of the length of coverage for autosomes

Alignment of latent periodicity region

```
6454977    tcttaaacatacaagcgatgaaattgag
6455005    aaaaagtaaactcgtaaattttccacca
6455033    aaaaacataaacccgtgatttttccacc
6455061    aaaaaatataaactcgtgatttttccgc
6455089    caaaaacgtaaacccgtgattttcccac
6455117    caaaaacgtaaactcgtgattttcccgt
6455145    caaaaacgtaaacccgtgaatttcccgc
6455173    caaaaacataaacccgtgattttcccgc
6455201    caaaaacgtaaatccgtaattttcccgc
6455229    caaaaacgtaaactcgtaattttcccac
6455257    caaaaaacgaaaacccgtgattttcccg
6455285    tcaaaaacgtaaacccgtgattttcccg
6455313    ccaaaaacataaatccgtgattttccca
6455341    ccaaaaacgtaaacccgtgattttcccg
6455369    tcaaaaacgtaaacccgtgaatttcccg
6455397    ccaaaaacataaacccgtgattttcccg
6455425    ccaaaaacgtaaatccgtaattttcccg
6455453    ccaaaaacgtaaactcgtaattttccca
6455481    ccaaaaaacgaaaacccgtgattttccc
6455509    gtcaaaaacgtaaacccgtgattttccc
6455537    gccaaaaacataaacccgtgattttctc
6455565    gccaaaaacgtaaatccgtgattttccc
6455593    gccaaaaacgtaaacccaagattttccc
6455621    gccaaaaacgtaaacccgtcattttccc
6455649    gctagaaacgtaaatccgtaattttccc
6455677    gtcaaaaccgtaaacctataatttttcg
6455705    ccaaaaacgtaaacccgtgattttccca
6455733    ccaaaaacgtaaacccgtaaaaagtgga
6455761    atccgtaaatattctaagtttga
```

```
6455005    aaaaa-g-taaactcgtaaattttcca-cca
6455033    aaaaa-cataaacccgtgattttcca-cca
6455062    aaaaa-tataaactcgtgattttccg-cca
6455091    aaaac-g-taaacccgtgattttccca-cca
6455119    aaaac-g-taaactcgtgattttcccg-tca
6455147    aaaac-g-taaacccgtgaatttcccg-cca
6455175    aaaac-a-taaacccgtgattttcccg-cca
6455203    aaaac-g-taaatccgtaattttcccg-cca
6455231    aaaac-g-taaactcgtaattttccca-cca
6455259    aaaaacg-aaaacccgtgattttcccg-tca
6455288    aaaac-g-taaacccgtgattttcccg-cca
6455316    aaaac-a-taaatccgtgattttccca-cca
6455344    aaaac-g-taaacccgtgattttcccg-tca
6455372    aaaac-g-taaacccgtgaatttcccg-cca
6455400    aaaac-a-taaacccgtgattttcccg-cca
6455428    aaaac-g-taaatccgtaattttcccg-cca
6455456    aaaac-g-taaactcgtaattttccca-cca
6455484    aaaaacg-aaaacccgtgattttcccg-tca
6455513    aaaac-g-taaacccgtgattttcccg-cca
6455541    aaaac-a-taaacccgtgattttctcg-cca
6455569    aaaac-g-taaatccgtgattttcccg-cca
6455597    aaaac-g-taaac-ccaagattttcccgcca
6455625    aaaac-g-taaacccgtcattttcccg-cta
6455653    gaaac-g-taaatccgtaattttcccg-tca
6455681    aaacc-g-taaacctataatttttcg--cca
6455708    aaaac-g-taaacccgtgattttcca-cca
6455736    aaaac-g-taaacccgtaaaaa

           AAAAC-G-TAAACCCGTGATTTTCCCG-CCA
```

Wraparound Dynamic Programming Algorithm for Tandem Repeat Alignment

G – is a weight matrix, α – score for gap opening,

β – score for gap extension, μ – score for substitution.

Initialization of row zero of G (*j* = 1,…, *L*): $G[0,0] = 0, G[0,j] = G[0,0] + \alpha + j\beta.$

Initialization of zero column (*i* = 1,…, *L*): $G[i,0] = G[0,0] + \alpha + i\beta$

For each row (*I* =1,…,*L)* the two passes are produced:

The first pass for the first column:

$$G[i,j] = \max \begin{cases} G[i-1,0] + \mu, \\ G[i-1,L] + \mu, \\ G[i-1,1] + \alpha + \beta, \\ G[i,0] + \alpha + \beta. \end{cases}$$

The first pass for other columns:

$$G[i,j] = \max \begin{cases} G[i-1,j-1] + \mu, \\ G[i,j-1] + \alpha + \beta, \\ G[i-1,j] + \alpha + \beta. \end{cases}$$

The second pass

$$G[i,j] = \max \begin{cases} G[i,j], \\ G[i,j-1] + \alpha + \beta. \end{cases}$$

Thanks for your attention!