SIMCOGEN: A database for collecting the location of similar regions in complete genomes.

Pierre Vincens[1,2], Cécile André[3], Anne Badel-Chagnon[1], Frédéric Guyon[1] and Serge Hazout[1]

(1) Equipe de Bioinformatique Génomique et Moléculaire, INSERM U436, Université Paris 7 - Denis Diderot, case 7113, 2 place Jussieu, 75251 Paris Cedex 05, France

(2) Département de Biologie (FR36), Ecole Normale Supérieure, 46 rue d'Ulm, 75230 Paris Cedex 05, France

(3) Département de Biomathématiques, CHU Pitié Salpêtrière, 91 boulevard de l'Hôpital, 75634 Paris Cedex 13, France

Goal

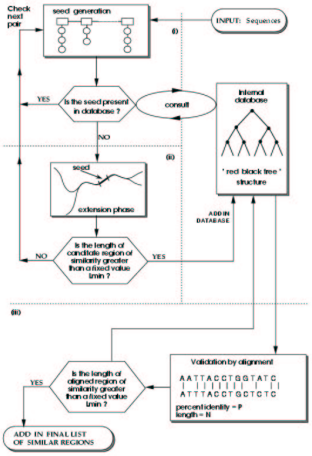
The knowledge of the location of duplicated regions into complete genomes is a major source of information to understand the mechanisms of their evolution [1, 4, 6, 8, 9, 13]. Our primary objective is to create a repository registering all pairs of similar regions found into a lot of complete genomes. Two regions are similar if the identity of their nucleotidic sequences is greater than S_{min} for a length greater than L_{min}.

Material and Methods

Data of SIMCOGEN

- Data are registered in a database using a relational model. The main tables include the location of the similar pairs and the informations describing the known regions of analysed chromosomes.
- The search of similar regions has been performed for a lot of chromosomes whose complete sequence is available. Are included procariotic (*B. subtilis*, *E. coli*, *M. genitalium*, ...) and eucaryotic organisms (*S. cerevisiae*, ...).
- The nucleotidic sequences and associated informations have been downloaded from:
  ftp://ftp.ncbi.nlm.nih.gov/genbank/genomes

Locating the similarities



The ASSIRC/DASSIRC [12, 11] approach is used for finding the similarities. The algorithm proceeds in three steps:

(I) Seed generation: A list of seeds is created. A seed is a pair of positions where the sequence has a common motif of fixed size k.

(II) Seed extension: The bounds of similar regions around the seed are determined. The method computes a 'random walk' [5] on each region from the seed and compares step by step these walks. The pairs of regions are recorded into an internal database by:
- . Using the Red-Black Tree structure (for rapid access)
- . Joining adjacent pairs together (to take gaps into account)
- . Checking for the inclusion of a seed in a previously recorded pair (to reduce the computation).

(III) Validation of pairs of similar regions: The similarity between two regions is checked using an alignment algorithm [7, 10]

Conclusions

SIMCOMGEN will make available the location of pairs of similar regions found in complete genomes. Recent works [2] have proved the interest of this repository to study the structure of subtelomeric regions of chromosomes of *Saccharomices cerevisiae*. However, it is required to complete the development of specific strategies as Mosaic approach [3] to analyse these data.

Availability

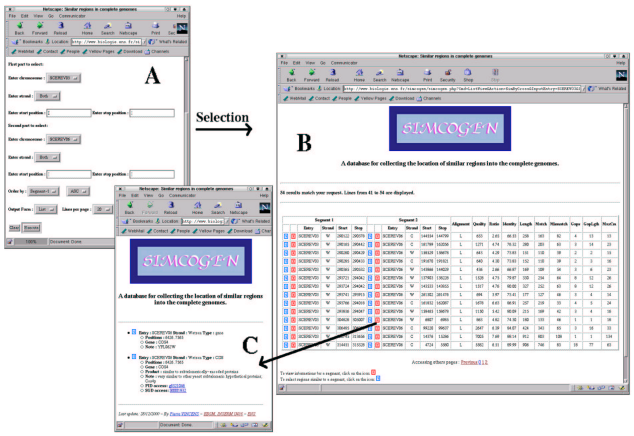
⇒ Interactive: http://www.biologie.ens.fr/simcogen

⇒ Flat data: ftp://ftp.biologie.ens.fr/pub/simcogen

Results

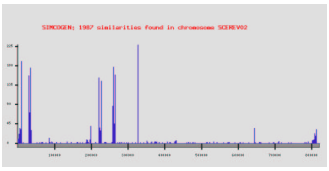
An example of interactive session

SIMCOGEN can be consulted using a web interface allowing to select and display the pairs of similar regions.

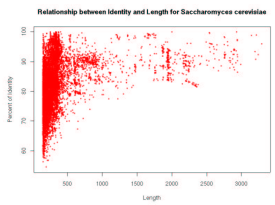


- A: Selection panel to choose the chromosome and the range of positions to consider. Many options allow the user to format the result.
- B: Result panel to view the pairs corresponding to the previous selection. For each segment of pairs, the I option gives complementary informations and the S option searches all similar pairs including this segment.
- C: Information panel to display data associated with a segment.

Graphic tools



Distribution of pairs of similar regions along the chromosome II of *S. cerevisiae*



Relationship between the identity and the length of similar pairs found in the genome of *S. cerevisiae*

Heterogeneity of the distribution of similar pairs in a lot of studied genomes

- The number of intrachromosomal duplications is variable:
  - . high in *Neisseria meningitidis*, *Deinococcus radiodurans chr 1*, *Ureaplasma urealyticum*
  - . low in *Pyrococcus abyssi*, *Pyrococcus horikoshii*, *Treponema pallidum*
- Very large duplications are observed for *Neisseria meningitidis* (**3171 bps**), *Saccharomyces cerevisiae (chr 2)* (**2864 bps**), *Treponema pallidum* (**1845 bps**), *Escherichia coli* (**1314 bps**).
- *Rickettsia prowazekii* is atypical: the pair of highest quality score has only a similarity of **64.2%**.

ID	*AC*	*Name*	*L*	*N*	*lms*	*pms*	*fms*
BSUBTI01	AL009126	*Bacillus subtilis*	4214814	176	3186	99.43	358
DRADUR01	AE000513	*Deinococcus radiodurans chr 1*	2648638	1273	2859	99.27	695
DRADUR02	AE001825	*Deinococcus radiodurans chr 2*	412348	42	600	99.27	695
ECOLIK01	U00096	*Escherichia coli*	4639221	637	3406	98.76	1314
MYPGEN01	L43967	*Mycoplasma genitalium*	580074	111	1244	91.68	117
NMENSB01	AE020098	*Neisseria meningitidis*	2272351	8971	3245	99.48	3171
PABYSS01	AL096836	*Pyrococcus abyssi*	1765118	51	1382	91.97	873
PYHORI01	Pyro-h	*Pyrococcus horikoshii*	1738505	75	1592	94.28	623
RPROWA01	AJ235269	*Rickettsia prowazekii*	1111523	244	328	64.20	26
SCEREV02	NC_001134	*Saccharomyces cerevisiae chr 2*	813140	1987	2927	99.93	2864
SCEREV04	NC_001136	*Saccharomyces cerevisiae chr 4*	1531929	205	2756	91.66	305
SCEREV12	NC_001144	*Saccharomyces cerevisiae chr 12*	1078172	156	2252	88.43	239
SCEREV16	NC_001148	*Saccharomyces cerevisiae chr 16*	948061	76	1365	98.68	270
TMARIT01	AE000512	*Thermotoga maritima*	1860725	132	1917	96.92	921
TPALLI01	AE000520	*Treponema pallidum*	1138011	53	1856	100.00	1845
UUREAL01	AF222894	*Ureaplasma urealyticum*	751719	1166	764	99.74	759
XFASTI01	AE003849	*Xylella fastidiosa*	2679306	148	2920	95.64	191

Similarities found for a lot of genomes: *ID*: SIMCOGEN identifier of the chromosome, *AC*: Accession number, *Name* : Current name of the organism, *L*: Length in nucleotide of the chromosome, *N*: Number of pairs of intrachromosomic regions of similarity found into the chromosome, *lms*: Length of aligned sequences for the pair of highest quality score, *pms*: Percent of identity of aligned sequences for the same pair, *fms*: Length of the larger identical segment between two similar sequences.

References

[1] G. Achaz, E. Coissac, A. Viari, and P. Netter, (2000), *Analysis of intrachromosomal duplications in yeast saccharomyces cerevisiae: a possible model for their origin*, Mol Biol Evol, 17, pp. 1268–75.

[2] C. André, *Etude des duplications et analyse de l'organisation subtélomérique dans les chromosomes de la levure Saccharomyces cerevisiae*, PhD thesis, Université Paris VII - Denis Diderot, (2000).

[3] C. André, P. Vincens, J. Boisvieux, and S. Hazout, (2001), *Mosaic: Segmenting multi-aligned dna sequences*, Bioinformatics, 17, pp. 196–197.

[4] E. Coissac, E. Maillier, and P. Netter, (1997), *A comparative study of duplications in bacteria and eukaryotes: the importance of telomeres*, Mol Biol Evol, 14, pp. 1062–1074.

[5] M. Gates, (1985), *Simpler dna sequence representations*, Nature, 316, pp. 219–219.

[6] B. Loftus, U. Kim, V. Sneddon, F. Kalush, R. Brandon, J. Fuhrmann, T. Mason, M. Crosby, M. Barnstead, L. Cronin, A. D. Mays, Y. Cao, R. Xu, H. Kang, S. Mitchell, E. Eichler, P. Harris, J. Venter, and M. Adams, (1999), *Genome duplications and other features in 12 mb of dna sequence from human chromosome 16p and 16q*, Genomics, 60, pp. 295–308.

[7] S. Needleman and C. Wunsch, (1970), *A general method applicable to the search for similarities in the amino acid sequence of two proteins*, Journal of Molecular Biology, 48, pp. 43–453.

[8] C. Seoighe and K. Wolfe, (1998), *Extent of genomic rearrangement after genome duplication in yeast*, Proceedings of the National Academy of Sciences of the United States of America, 95, pp. 4447–4452.

[9] ——, (1999), *Updated map of duplicated regions in the yeast genome*, Gene, 238, pp. 253–61.

[10] T. Smith and M. Waterman, (1981), *Comparison of biosequences.*, Advance in Applied Mathematics, 2, pp. 482–489.

[11] P. Vincens, A. Badel-Chagnon, F. Guyon, C. André, and S. Hazout, *Stratégie de recherche des similitudes intragénomiques*, in JOBIM 2000, G. Caraux, O. Gascuel, and M. F. Sagot, eds., AGRO Montpellier - LIRM, Montpellier, (2000), pp. 383–389.

[12] P. Vincens, L. Buffat, C. André, J. Chevrolat, J. Boisvieux, and S. Hazout, (1998), *A strategy for finding regions of similarity in complete genome sequences*, Bioinformatics, 14, pp. 715–725.

[13] K. Wolfe and D. Shields, (1997), *Molecular evidence for an ancient duplication of the entire yeast genome*, Nature, 387, pp. 708–713.