Assemblage de novo distribué de génomes avec Ray et RayPlatform: des graphes de Bruijn aux polytopes

Sébastien Boisvert

Vendredi 1er mars 2013, Heure: 12h00, Local: PLT-3775
Séminaires du département d'informatique et de génie logiciel
Université Laval

- Directeur: Jacques Corbeil
- Codirecteur: François Laviolette

- Projets financés par les Instituts de recherche en santé du Canada
- Temps de calcul avec Calcul Canada (Calcul Québec et SciNet), et accès au Blue Gene/Q à SOSCIP

Blocs de la vie

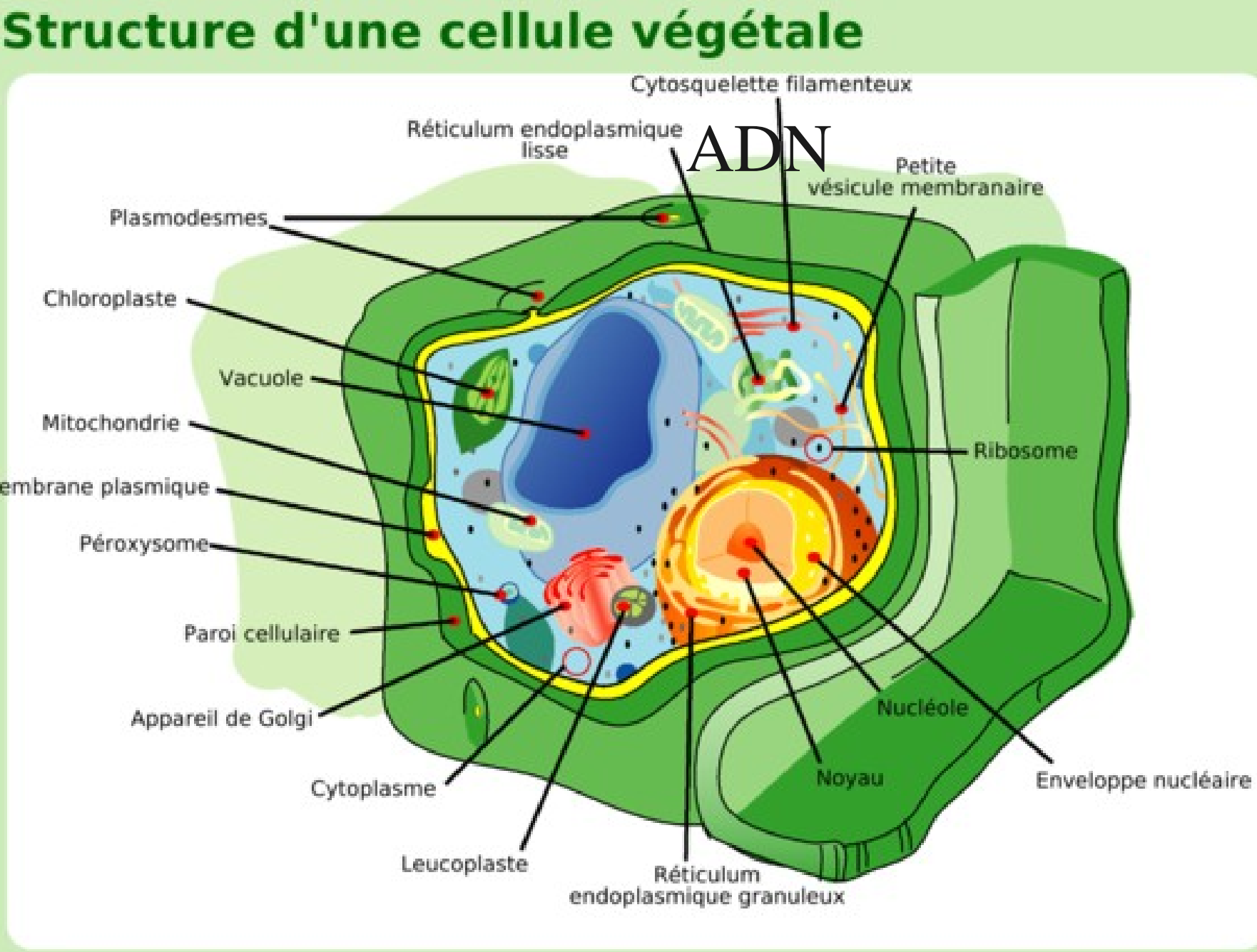


Image: http://fr.wikipedia.org/wiki/Cellule_%28biologie%29

Information biologique

ADN

ARN

protéine

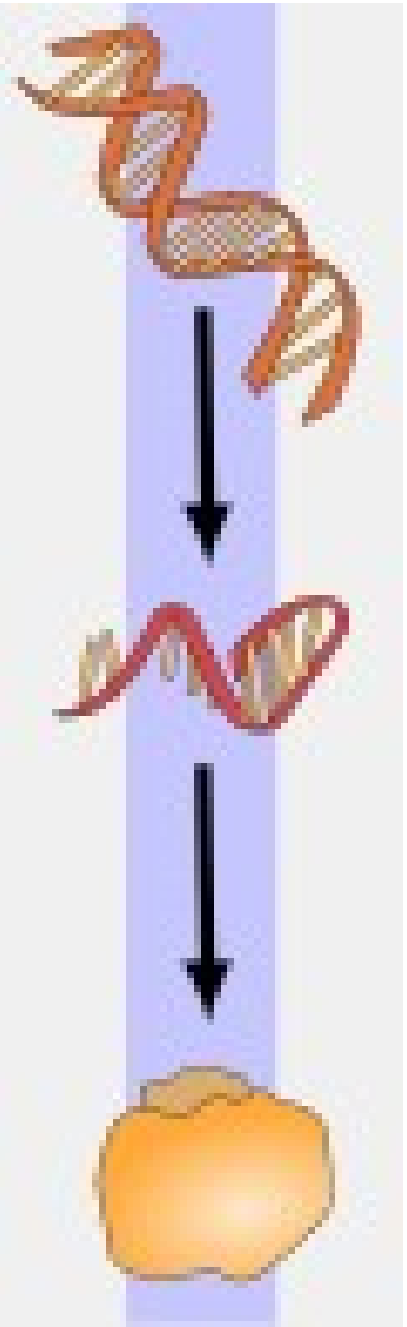


Image: http://fr.wikipedia.org/wiki/Dogme_central_de_la_biologie_mol%C3%A9culaire

Génome

- Étymologie: gène et -ome
- Déf.: ensemble du matériel génétique

Lire l'ADN

- Détecter les agents infectieux
- Expliquer des conditions
- Étudier des systèmes

Écrire l'ADN

- Corriger des défauts dans l'ADN

Séquençeurs d'ADN

Tubes contenant de l'ADN

Fichiers contenant de l'ADN

Image: http://www.dnavision.com/images/HiSeq2000.jpg

Deux types d'analyse

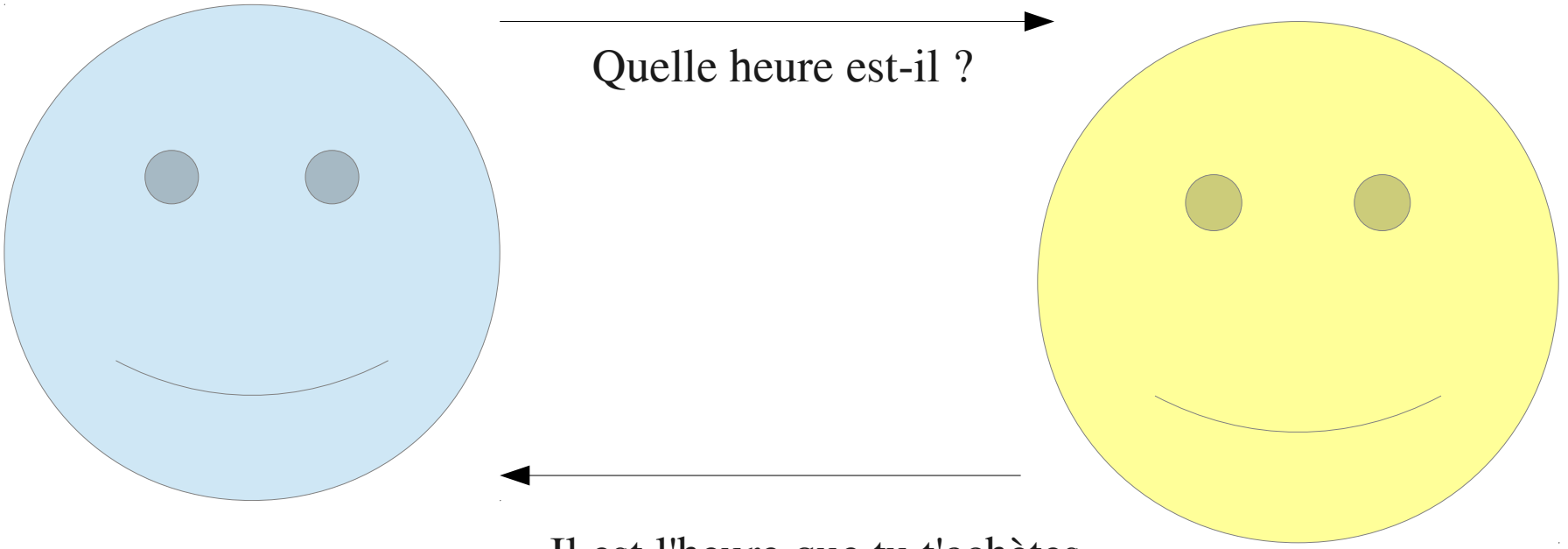
- Comparer l'ADN lu à une référence
- Assembler l'ADN lu sur lui-même (casse-tête)

Flicek & Birney

Nature Methods 2009

http://www.nature.com/nmeth/journal/v6/n11s/full/nmeth.1376.html

Passage de message



Granularité

- Granularité grossière
- Granularité fine

Architecture versatile

- def progresser():
  - while doitContinuer():
    - recevoirLesMessages();
    - traiterLesMessages();
    - faireLesCalculs();
    - envoyerLesMessages();

MPI (“Message Passing Interface”)

- Interface de passage de message
- Opérations collectives
- Opérations non-collectives

MPI 3.0, Chapitre 2

- MPI_Init
- MPI_Finalize

MPI 3.0, Chapitre 3

- MPI_Isend
- MPI_Iprobe
- MPI_Recv
- MPI_Test

MPI 3.0, Chapitre 13

- MPI_File_open
- MPI_File_set_view
- MPI_File_write
- MPI_File_read
- MPI_File_close

Indexer l'ADN

- Le graphe de Bruijn
- Entier k
- Alphabet A={“A”, “T”, “C”, “G”} (pour l'ADN)
- Sommets: tous les mots de A^k
- Arêtes: |A| parents et |A| enfants pas sommet
- Exemple pour les enfants:
  A**TGTAC** -> **TGTAC**A
  A**TGTAC** -> **TGTAC**T
  A**TGTAC** -> **TGTAC**G
  A**TGTAC** -> **TGTAC**C

- |A|^k sommets et |A|^(k+1) arêtes
- Diamètre de k

Compeau, Pevzner & Tesler
Nature Biotechnology 2011
http://www.nature.com/nbt/journal/v29/n11/full/nbt.2023.html

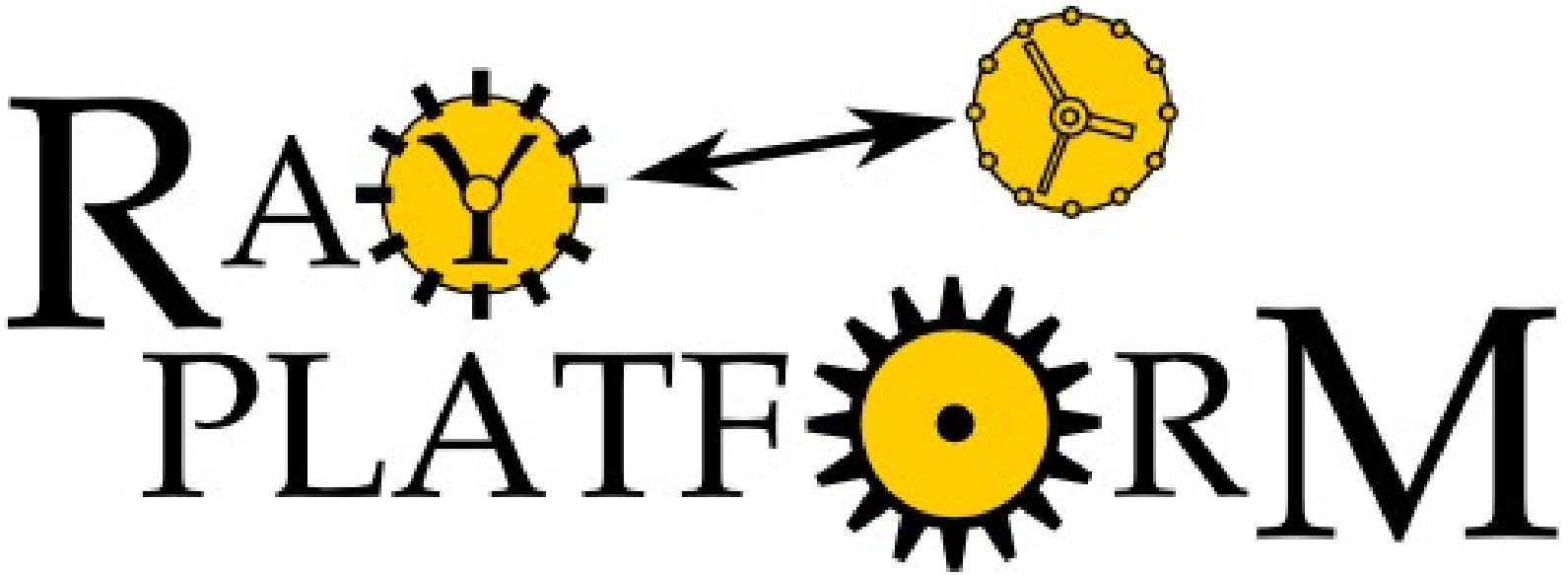
Sous-graphe du graphe de Bruijn

- Pour |A| = 4 et k = 31
- 4611686018427387904 sommets
- 18446744073709551616 arêtes

Solution

- Sous-graphe généré avec les données
- Un génome de 4 600 000 aura 2 * 4 600 000 sommets dans le pire cas (sans répétitions)

Ray Platform



C++ 1998

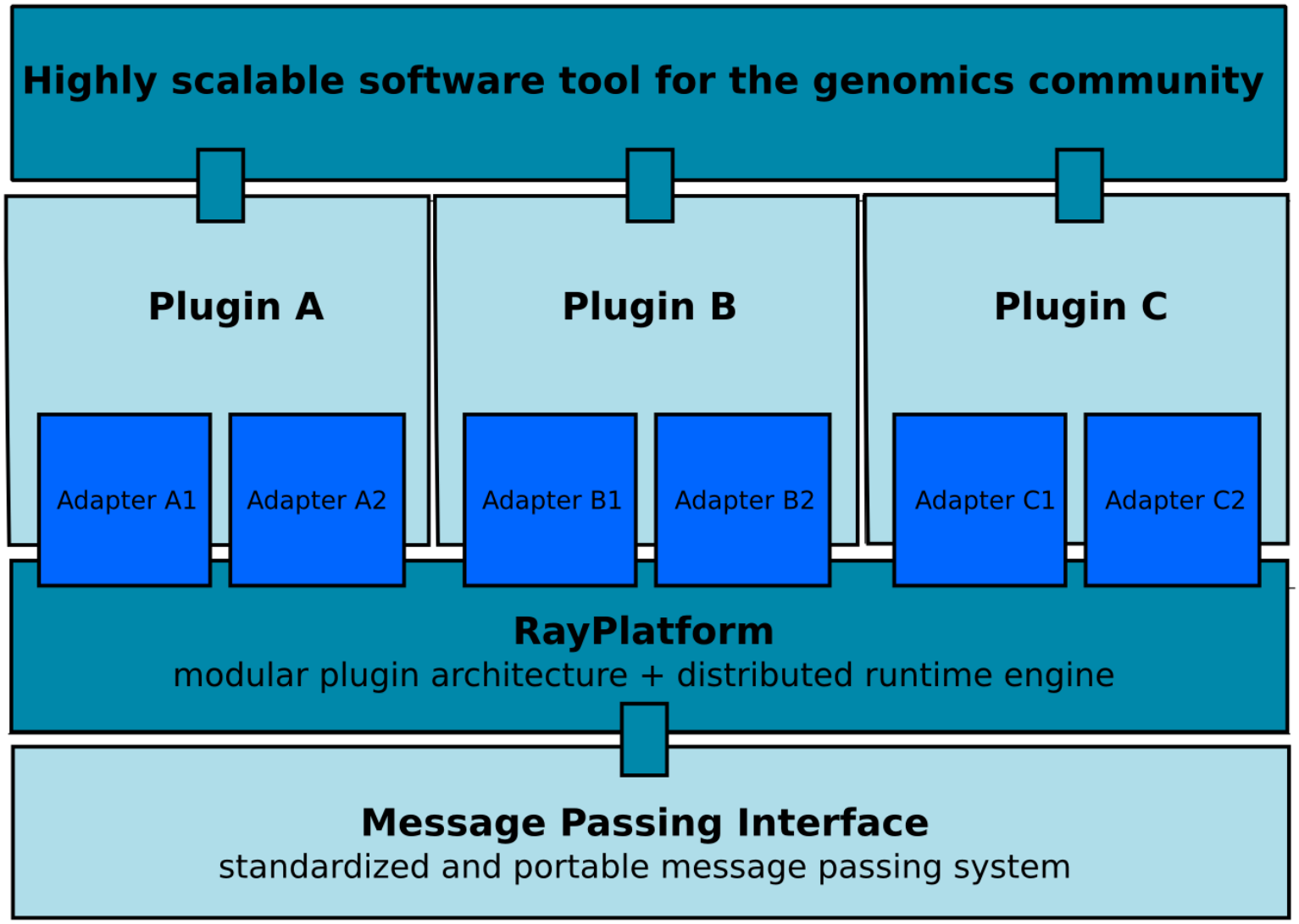
MPI 2.0

LGPLv3

Godzaridis et collaborateurs

En édition (rédaction complétée) pour Big Data

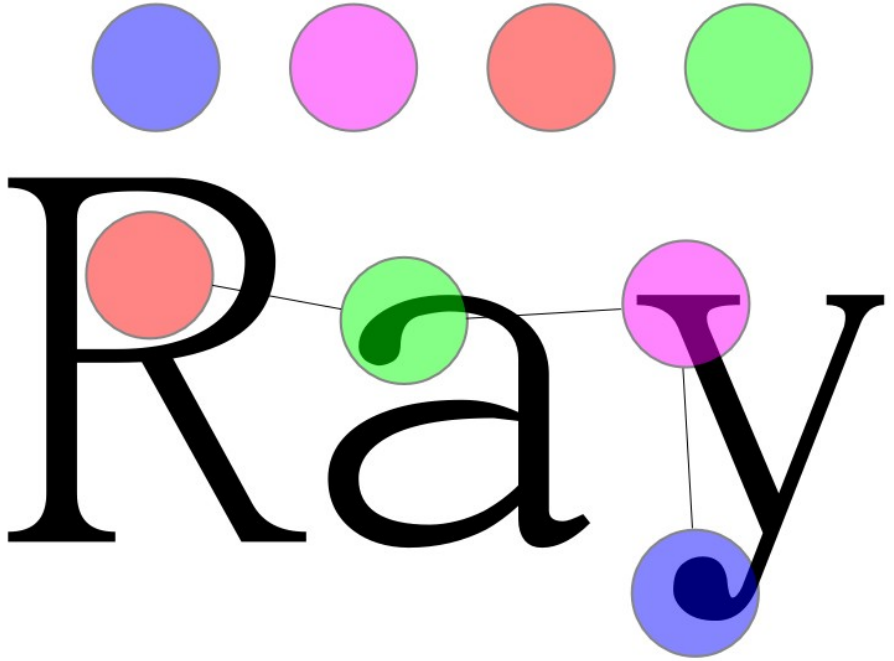
Architecture de Ray Platform



Quick facts:

- RayPlatform assists developers during the content creation
- A software tool for the genomics community is implemented as plugins
- Plugins are executed by a distributed runtime engine
- Communication is portable thanks to MPI

Construire des cartes pour les génomes



Boisvert, Laviolette & Corbeil

Journal of Computational Biology 2010

http://online.liebertpub.com/doi/abs/10.1089/cmb.2009.0238

Pour des mélanges de génomes

- Plusieurs génomes
- Abondances selon une loi de puissance

Boisvert, Raymond, Godzaridis, Laviolette & Corbeil
Genome Biology 2012
http://genomebiology.com/2012/13/12/R122/abstract

Assemblathon 2

- Bradman et collaborateurs (91 auteurs)
  en révision chez GigaScience
  http://gigasciencejournal.com/

  Liens:
  http://arxiv.org/abs/1301.5406
  http://assemblathon.org/?tag=assemblathon2

Polytope

- Base B=8
- Dimension d=2
- Alphabet A = {0,1,2,3,4,5,6,7}
- Sommets (x,y) dans A x A
- B^d sommets
- Degré k=d * (B-1)

Routage adaptatif avec un polytope

- Utiliser les sommets et les arêtes

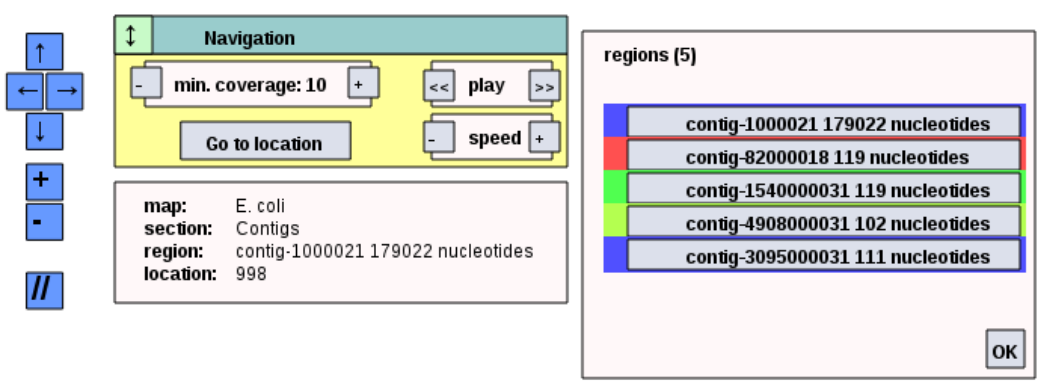
Routes dynamiques

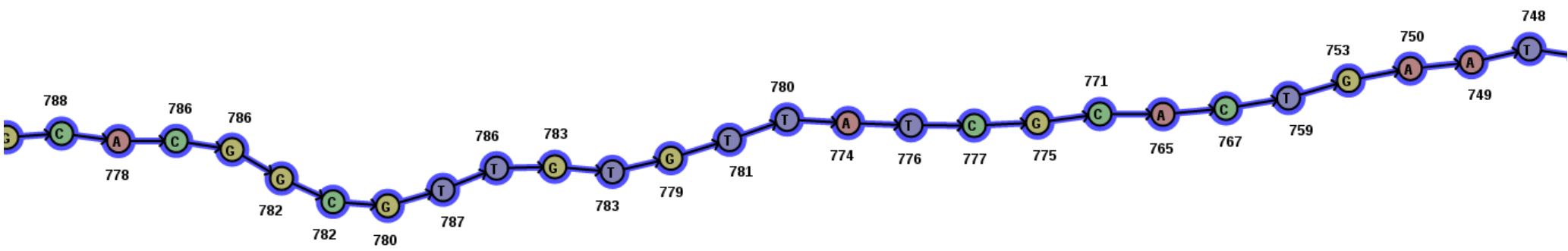
- (x,y) vers (x,y): <(x,y)> (0 saut)
- (x,y) vers (x,w): <(x,y), (x,w)> (1 saut)
- (x,y) vers (w,y): <(x,y), (w,y)> (1 saut)
- (x,y) vers (w,z): <(x,y), (w,y), (w,z)> (2 sauts)
- (x,y) vers (w,z): <(x,y), (x,z), (w,z)> (2 sauts)

Algorithme simple

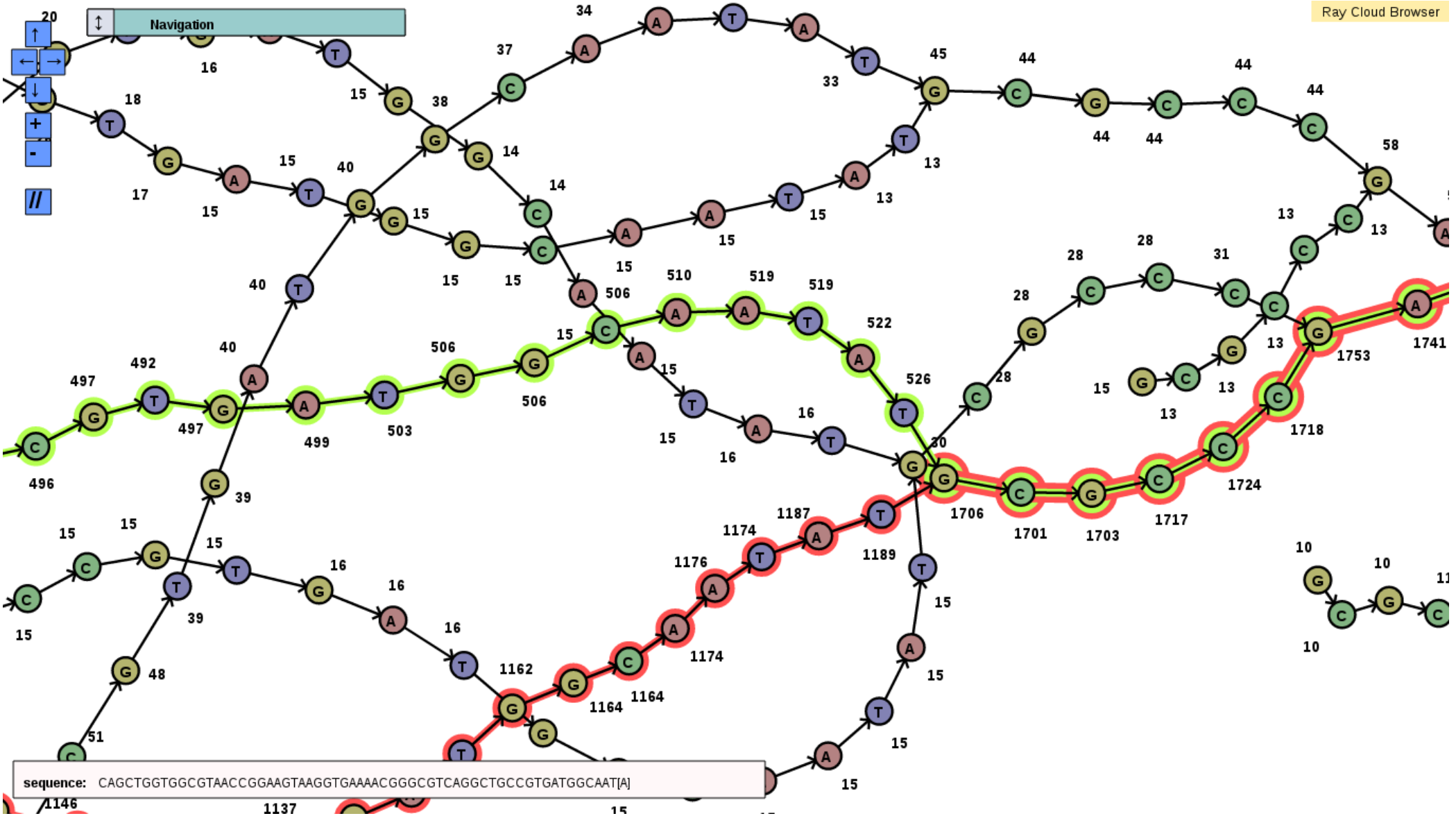
- Compteur du nombre d'opérations sur chaque arête
- Sélection de l'arête la moins utilisée

Visualiser l'ADN à la Google Maps

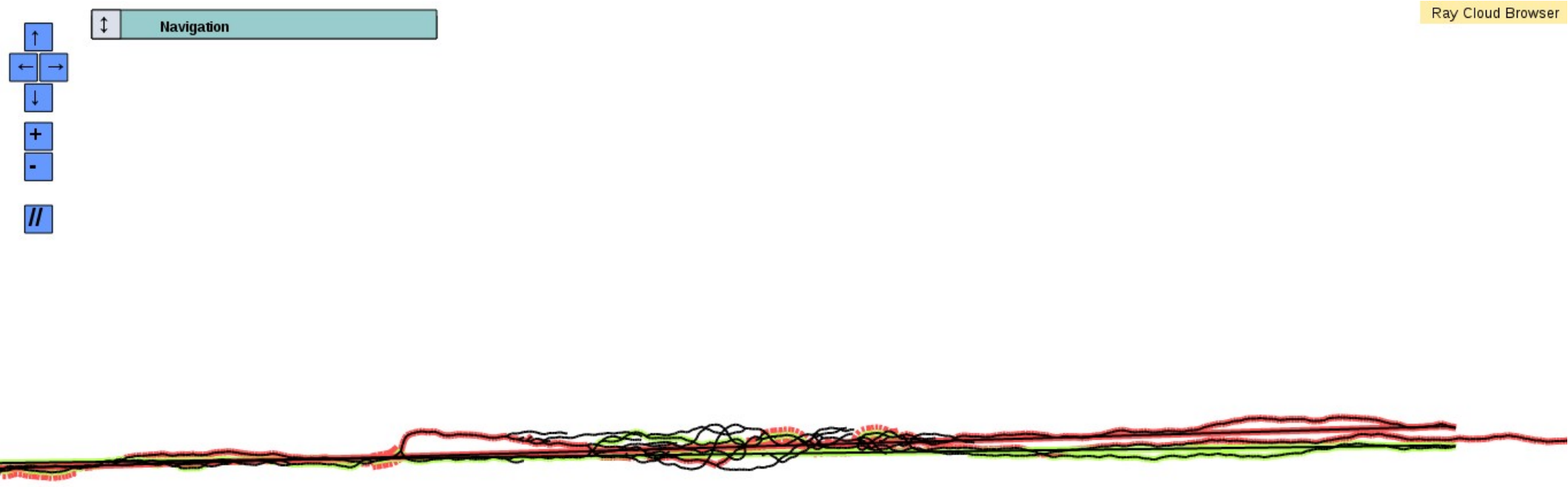




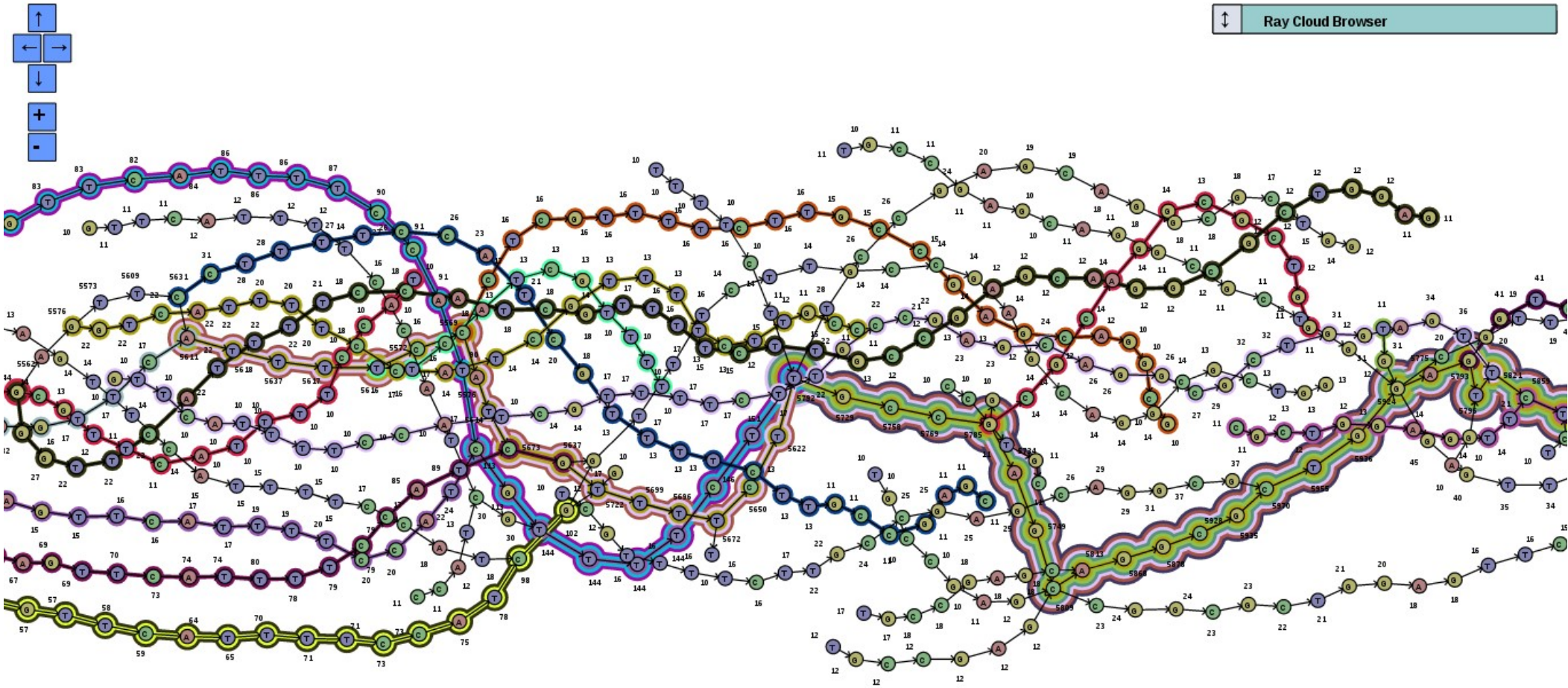
La carte de métro génomique



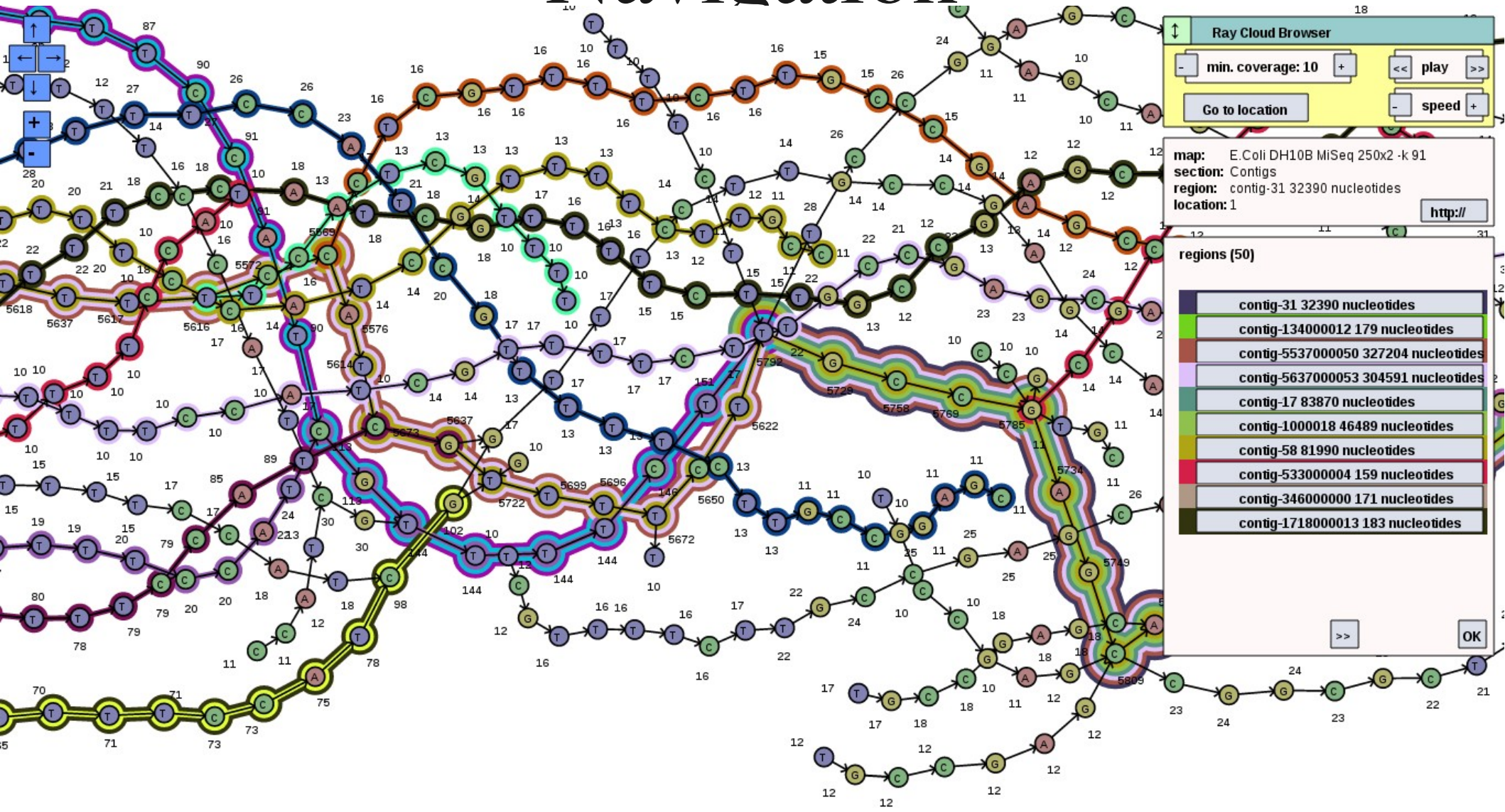
À vol d'oiseau



Carte de métro génomique



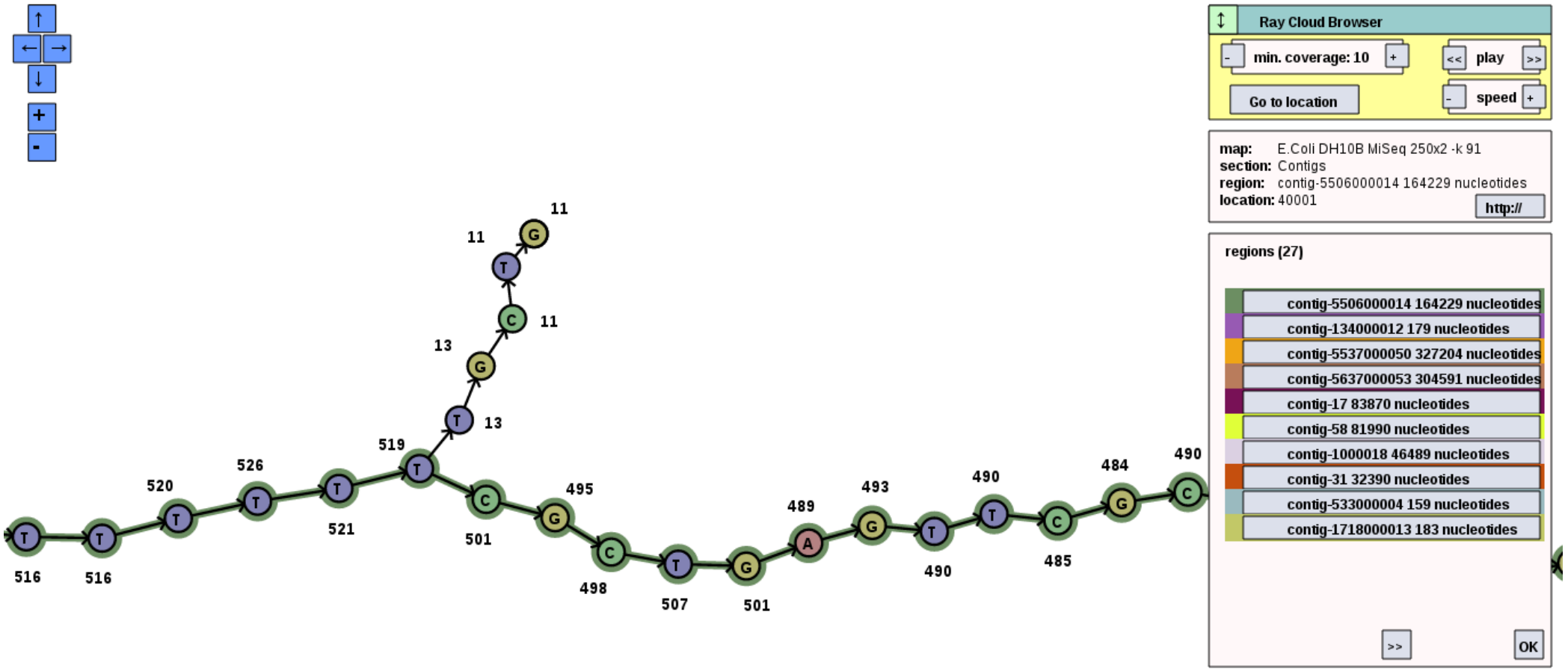
Navigation



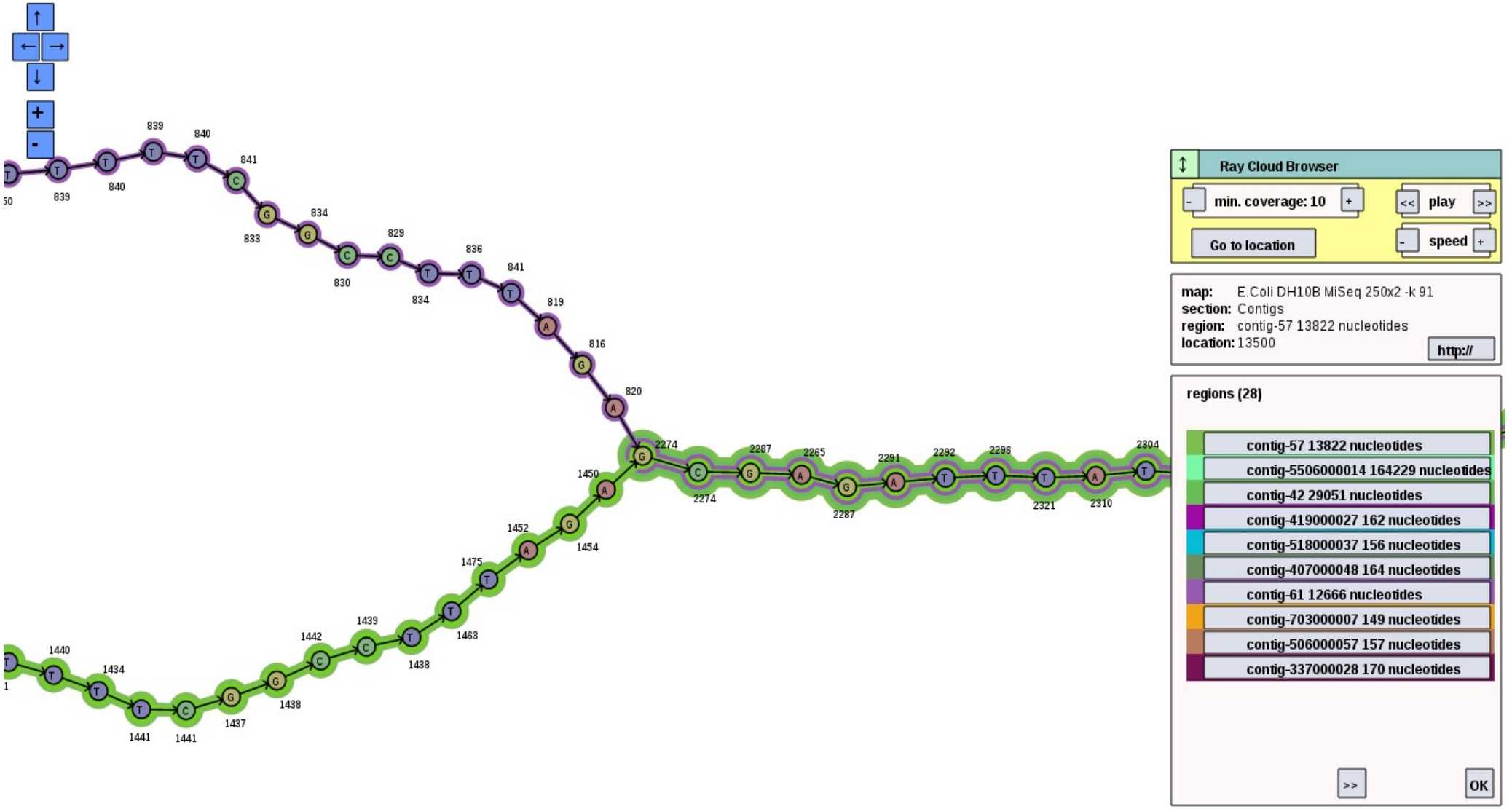
Pour y aller:

http://browser.cloud.boisvert.info/client/?map=0&section=3&region=41&location=0

Région unique



Région répété



Questions

- (Je dois partir vers 13:00.)