Réunion PhyloAlps #1

2017/02/20

Participants

- Roland Douzet
- Christophe Perrier
- Eric Coissac
- Frédéric Boyer
- Sébastien Lavergne
- Anthony Hombiat

Considérations générales

- BD **séquences ADN des plantes de l'arc Alpin**
- Base de données pour la réduction du génome :
  - **Genome skimming**
  - RADseq : sélection aléatoire d'un sous ensemble du génome (reproductible)
  - Capture d'exon : cible spécifiquement les protéines
- BD ouverte à tous les organismes experts du domaine
- Répertorie les plantes angiospermes et vasculaires

> Au-delà du problème technique, problème politique : les instituts sont crispés sur leur données : il faut les valoriser via la BD, externaliser la responsabilité et rendre les jeux de données des différentes institutions modulaires pour pouvoir les gérer indépendamment.

Données disponibles dans BD existantes

- **Herbier PhyloAlps** (en cours de réalisation), QR code (référence espèce et séquence)
- **Herbier du CBNA** : conservatoire Botanique Nationaux (CBN Alpin à Briançon et CBN Méditerranéen)
- **GBIF** : Global Biodiversity Information Facility (occurrences, observations, /!\ couverture)
- **Androsace** : traits biologiques de référence pour les plantes Alpines (voir avec Julien pour la disponibilité des données)
- **IFB** : Institut Français pour la Bioinformatique
- **FRB** : Fondation pour la Recherche en Biodiversité
- **BOLD** : pas le génome complet (marqueurs taxonomiques, 2 gènes matK et rbcL)
- **INSD** : International Nucleotide Sequence Database Collaboration :
  - **NCBI GenBank** : National Center for Biotechnology Information
  - **EMBL** : European Molecular Biology Laboratory
  - **DDBJ** : DNA DB of Japan European Molecular Biology Lab
- **UniProtKB** (Swiss-Prot) utilise EMBL
- **Phylota** : browser pour les plantes
- **FloraAlpina** (version électronique ?)
- **Tela Botanica** : association de botanique de référence pour la botanique numérique
- **Plant list** (cf. Kew project : séquençage d'exons)

Interface de restitution

- Recherche génomique type sur la BD :
  - Synonymie sur le référentiel taxonomique
  - Entonnoir : Taxon > librairie > échantillon

- Pour chaque collection :
  - Description
  - Emprise fonctionnelle
  - Emprise géographique
- Une page par taxon ? par biome ?
  - TaxId
  - Binôme genre-espèce et auteur (cf. Carl Von Linné)
  - Photos et scans d'herbier (serveurs d'images ? modèle d'herbier numériques existants ?)
  - Séquences d'ADN
  - Aires de répartition
  - Plotlist
  - Traits biologiques
  - Génome skimming
  - Accès part d'herbier
  - Silicathèque (conservation supérieure de l'ADN grâce à une sécheresse maximum)
  - Localisation GPS (souplesse, pas obligatoire)
  - Taxon coverage
  - Lien vers GBIF France
  - Critères remarquables
  - Chaque échantillon est rattaché à
    - une collection primaire
    - 0 ou plusieur(s) collection(s) secondaire(s)

Qualité des données

- 2 types de données :
  - Données pérennes (séquence, scan) -> entrées une fois pour toutes
  - Données périssables (échantillons physiques) -> méthodes et outils pour la mise à jour
    - Que faire des calculs longs qui s'exécutent alors que les données en entrée changent ?
- Modération, pas de modification directe de la BD
- Utilisateur identifié en tant que référent et garant :
  - Qui ?
  - Quel labo ?
- Référentiel mondial des noms d'auteur
- Utilisateur responsabilisé, rattaché à une collection
  - batchs de soumission
- Modèle de qualité de la donnée multi-critères
  - Confiance en l'auteur
  - Avis subjectif de l'auteur
  - Complétude des caractéristiques générales
  - Complétude des caractéristiques génomiques
  - Avis des utilisateurs
    - Système d'annotations : cf. features table EMBL/GenBank doc (système de preuves "evidences")
    - Issue tracking system (cf. Alain Viari, DR INRIA, bioinformaticien Herbs)
- Processus de validation de la données en plusieurs étapes
  - Vérification syntactique et sémantique (automatique)
  - Identification des outliers (semi-automatique)
- Elaboration de méthodes pour l'exploitation conjointe de référentiels taxonomiques hétérogènes :
  - NCBI
  - Plant List
- Suivi de versions
  - Modifications en continu ?
  - Nouvelle mouture tous les X mois ?