Biostrings directions

Improvements to DNAStringSet

CircularDNAString

Improvements to PDict/matchPDict()

matchLRPDict()

Biostrings C interface

Improvements to PDict/matchPDict()

Support for IUPAC ambiguities in the reads

Let the user limit the number of matches per read

Support preprocessing of multiple lanes together

Support for indels

Harris' suggestion: Treat a 'max.mismatch' value that is strictly between 0 and 1 as an error rate (so that the actual max number of mismatches auto-adjust to the length of each pattern)

Patrick's suggestion: Give the user the option to make matchPDict() return directly the coverage of the hits

6 lanes / 4.5 M 35-mers per lane

Mapping 1 lane at a time

Preprocessing:

- Direct reads + reverse complements = 2 x 4.5 = 9 M reads to preprocess
- Preprocessing time < 1 min.
- Size of the resulting PDict object: 2.64 GB

Walking the Mouse genome: 31 min. (exact matching)

Total time for the 6 lanes: 6 x (1 min. + 31 min.) = **3 h 12 min.**

Mapping 6 lanes at a time

Preprocessing:

- Direct reads + reverse complements = 54 M reads to preprocess
- Preprocessing time = 9 min.
- Size of the resulting PDict object: 12.26 GB

Walking the Mouse genome: **1 h** (exact matching)

Timings obtained on a 64-bit openSUSE 10.3 system with 64 GB of RAM

ReadMatcher package (coming soon)

For mapping HTS reads against a reference genome

Will build on top of matchPDict() and the BSgenome infrastructure to map reads against a whole genome (or parts of it). With a focus to do things more in a MAQ/Bowtie fashion (i.e. high level tools, no need for the user to know the details/tricks of the PDict/matchPDict tool, with output written to a file, etc...)

Will also provide tools for more specific types of mappings. E.g. an RNA matcher (allows each read to match with 1 gap of arbitrary length in it). By reusing the Biostrings C level infrastructure -> little C code is required (the complexity is encapsulated on the Biostrings side)

Pedagogical purpose: will be a good illustration of how to reuse the PDict/matchPDict infrastructure + the Biostrings C interface

Long term goal: make it easier for people to write specific mappers to target specific problems