作成:2013 年 5 月 2 日

最終更新:2014 年 1 月 29 日

ゲノム情報と電子化医療情報等の統合によるゲノムコホート研究の推進

―大規模分子疫学コホート研究の推進と統合

次世代シークエンサーデータ解析担当バイオインフォマティシャン育成カリキュラム
-実験出身者-

独立行政法人 国立がん研究センター

0. このカリキュラムの位置付け

人材育成のカリキュラムは可能な限り個別化されねばならない。この文書は 2013 年の時点で本研究において実際に人材育成を試みた、主に実験研究を経験してきた対象者を想定しており、そのまま将来にわたって万人に一般化できるものではない。

1. 目的と目標

1.1. 最終目的

Next-generation sequencing (NGS)を中心としたゲノムデータの急増に伴い、これらのデータと臨床情報とを結び付け、生物学的な意味付けを行うことが出来る人材不足が顕著となった。本事業では NGS のデータ解析が行える人材を育成することを最終目標とする。

1.2. 育成期間と目標

NGS データ解析を有効且つ効率的に行うには、

A) PC クラスター計算機上でのジョブの効率的な実行

B) 使用するプログラムの性質の把握(アルゴリズム、ボトルネック、使用メモリ等)

C) 必要なツールの作成

D) 分子生物学(ゲノム)の知見に基づいた解析計画

を行う能力が必要である。実際のデータ解析では、計算機センターの様な施設を利用することもあれば、自分で計算機システムをデザイン、導入して使う必要があることもある。そのため計算機のハードウエアに対する知識も一定程度必要になる。

上記の知見を備えた人材を育成するには、

1年目　計算機の取り扱い方及び統計学に対する知識を身に付け、チップによるオミックスデータの解析を行う。

2年目　PC クラスターを利用した NGS データ解析を行う。

3年目　必要な解析を自身で立案し、一連の解析を実施する。

の様な計画が考えられるが、育成対象者の事前知識により育成計画は大きく異なる。

2. 1年目の計画

育成対象者を医学、生物学系の学部教育を受けた者とし、バイオインフォマティクスに必要な数学、情報学の習得から行う。そのため初めの半年間は座学を中心とするが、全てを一度に身に付けるのは困難であるので、まずは関連書籍においてどこに何が書いてあるのかを把握し、使いながら覚えていくための準備期間とする。具体的には、

① Linux の使い方の習得(実用上ほとんどの作業は Linux 上で行うことが要求される)。
② プログラミング言語 C 第 2 版 (共立出版. カーニハン, リッチー著)の通読
③ プログラミングに活かすデータ構造とアルゴリズムの基礎知識 (アスキー,今泉貴史著)の通読
④ 数値計算のつぼ (共立出版. 二宮 市三 編) の通読
⑤ バイオインフォマティクスのためのアルゴリズム入門 (共立出版, Pevzner, Jones 著)
⑥ CLAPACK の SVD による主成分分析の実装 (make や gdb、valgrind の使い方も学ぶ)
⑦ ロベールの C++入門講座 (毎日コミュニケーションズ, ロベール著) の通読
⑧ 初めての Perl 第 6 版 (オライリージャパン, 近藤 嘉雪 翻訳) の通読
⑨ 統計学 (学部程度)の習得

を予定している。ただし「バイオインフォマティクスのためのアルゴリズム入門」は希望者を募り輪講にする予定。主成分分析(ケモメトリックス―化学パターン認識と多変量解析; 共立出版 佐々木慎一・宮下芳勝 著 参照)とは結局は行列の対角化であるが、必要であれば線形代数の勉強も独自に行う。主成分分析の入力はチップによるオミックスデータ(縦にサンプルが並び、行毎にプローブの測定値がタブ区切り)を想定している。

この後に実際のオミックスチップデータの quality control 解析、臨床情報との相関、関連解析を行うが、その際に

⑩ 生物統計
⑪ R
⑫ sh script

も使いながら学ぶ。また必要に応じて pLaTeX2ε も習得する。実践的な生物統計の教科書として、

A) 「ロジスティック回帰分析―SAS を利用した統計解析の実際」 (朝倉書店) 丹後 俊郎, 高木 晴良, 山岡 和枝
B) 「Statistical Methods in Medical Research」(Wiley-Blackwell) Peter Armitage, Geoffrey Berry, J. N. S. Matthews
C) 「Statistical Methods in Cancer Research Vol.1: The Analysis of Case-Control Studies」(IARC) N. E. Breslow and N. E. Day
D) 「Modern Epidemiology」 (Lippincott Williams & Wilkins) Kenneth J. Rothman, Timothy L. Lash Associate Professor, Sander Greenland

を挙げておく。

2.1. Linux で良く使うコマンド

UNIX/Linux では小さなツールを組み合わせてやりたいことを実現する。組み合わせる方法として、標準入力、標準出力が定義されており、これらをパイプで繋ぐことにより柔軟に目的が達成出来る仕組みになっている。そのためある程度の知識があると非常に強力であるが、コマンド類を一々覚えなければならない。以下、表 1 及び表 2 に使用頻度の高いコマンドを列挙した。尚、C-c の「C-」とは「control キーを押しながら」を意味する。

表 1 基本ツール:やや大物

bash (intractive)	C-c	C-z	bg	fg	↑	↓	tab 補完
	alias	.bashrc	環境変数				
manual	man	info	google？				
emacs	必要に応じて caps キーと control キーを入れ替える。						
	C-x C-s	C-x C-c	C-x C-w	C-x C-f	C-x k	C-x u	C-g
	C-p	C-n	C-f	C-b	C-a	C-e	C-l
	C-x b	C-x 2	C-x o	C-x 0	C-x 5 2	C-x 5 0	
	C-space	C-w	C-k	C-d	C-y	C-x r k	C-x r y
sed	標準出力	標準入力	|	1>、2>	s///		
X Window system	xhost	DISPLAY					

表 2 基本ツール:コマンド類

ファイル操作	ls	cp	mv	rm	ln	mkdir	less
	head	tail	cat	sort	uniq	wc	diff
	cut	paste	chmod	touch	unix2dos	dos2unix	nkf
検索	grep	find	locate				
圧縮、アーカイブ	gzip	bzip2	zip	7z	tar		
システム、プロセス	df	du	w	top	ps	pwd	
確認	jobs	kill	killall				
ネットワーク通信	ssh	scp	host	ping			
開発環境	gcc	gdb	valgrind	make			

3. 2 年目

実際に NGS データのデータ解析を行う。1 年目の補完を含め、NGS データ解析で必要となる知識として、

① データマイニング

② 遺伝統計学(オプション)

③　Linux 管理
④　並列化(MPI、スレッド)
⑤　バッチシステムの利用
⑥　Java (オプション)
を習得する。参考図書として、
A)　「パターン認識と機械学習(上/下)」（丸善出版）　C. M. ビショップ
B)　「遺伝統計学入門」（岩波書店）鎌谷直之
C)　「ゲノム 第 3 版」（メディカルサイエンスインターナショナル) T. A. Brown
を挙げておく。

4.　3 年目

論文やマニュアル等を参考にアルゴリズムを評価し、必要なソフトウエアをインストールして実行し、この結果を相互に比較することにより、目的に応じたパイプラインを構築する。また必要であれば数値計算ソルバーの開発も行う。

5.　付録

課題を実施する上での具体的な指示例を示す。

5.1.　C 言語課題:LAPACK を用いた SVD による PCA

目的:

数値計算ライブラリーを使ったプログラミング方法を学び、gdb によるデバッグ、valgrind によるメモリリークの検出方法を学ぶ。

1. ケモメトリックス―化学パターン認識と多変量解析」を初めから 55 ページまでを読む。
2. 「ニューメリカルレシピ・イン・シー　日本語版」を購入し、「2.9 特異値分解」(p.73-84)までを読む。付属のプログラムを理解する必要ない。
3. 自分の Linux マシンに BLAS 及び LAPACK をインストールする。
   http://www.netlib.org/lapack/
   http://www.netlib.org/lapack/faq.html
4. 入力データとして、mRNA 発現を測定するマイクロアレイを想定し、LAPAC の dgesdd (http://www.netlib.org/lapack/lug/node32.html)で SVD による PCA を実装する。入力データは、3 万プローブ×100 検体程度と考えて、自分で模擬データ(乱数で良い)を作製する。
5. 結果が正しいかどうかをどの様にチェックすれば良いのか考え、実行する。

5.2. C++課題: memx への Cauchy 分布による解析機能の追加

目的:

C++によるオブジェクト指向プログラミングの実例に触れ、Run-Time Type Information によるインターフェイスと実装の分離の 1 例を見る。また make や gprof についても経験する。

1. memx のドキュメント及びソースコードからプログラム構成を把握する(方法論を理解する必要は無い)。

2. memparameters に、model=normal or model=cauchy を追加し、MemParameter クラスに追加した string model 変数に格納させる。

3. MemCauchyPixels クラスを作成し、memPara.model 変数によって Cauchy 分布による解析も選択出来る様にする。

5.3. 統計学及び perl 課題

目的:

中心極限定理の適用範囲を確認し、perl を使って実際に確かめる。

1. 独立に正規分布に従う確率変数の比は、Cauchy 分布に従うことを示せ(x, y～ N(0,1)の時、x/y は Cauchy 分布になる)。

2. 正規分布及び Cauchy 分布からランダムサンプリングを行い、平均値の分布をヒストグラムで示せ。例えば、100 回サンプリングして平均値を計算するプロセスを 1000 回行って、平均値のヒストグラムを作る。一様乱数から Cauchy 分布に従う乱数を作成する方法は、

http://www.ntrand.com/jp/cauchy-distribution/

等を参照。

3. オミックスデータはデータの規格化の際に何らかのデータの和で割られることも多い。この場合近似的には独立に正規分布に従う統計量(ただし最低値が存在することが多い)の比(Hinkley 分布)を取り扱うことになる。この時 t 検定及び Wilcoxon の順位和検定を行うことの是非を考えよ(必要に応じて独立に Cauchy 分布に従う 2 群間の比較を t 検定及び Wilcoxon の順位和検定で行い、この p-p plot を作成)。

5.4. オミックスデータ解析の課題

メチレーションチップデータの quality control 解析を再現せよ。