生物情報学用ライブラリ BioRubyの歩み

後藤　直久
ngoto@gen-info.osaka-u.ac.jp
大阪大学微生物病研究所附属遺伝情報実験センター

自己紹介

- 名前：後藤直久（ごとうなおひさ）
- 大阪大学微生物病研究所附属遺伝情報実験センター　ゲノム情報解析分野　助教
- 専門:
  - バイオインフォマティクス
  - ゲノム情報解析
  - ゲノムの進化

本日の内容

- BioRubyとは？
- Google Summer of Code

BioRubyとは？

- 生物情報学（バイオインフォマティクス）用ライブラリ
  - 生物学と情報科学が融合した学問
    - バイオ=生物学
    - インフォマティクス＝情報科学
    - バイオ＋インフォマティクス＝バイオインフォマティクス
    - 生物の情報を解析することによって、生命現象を解明
- フリーソフトウェア
- http://bioruby.org/
- http://github.com/bioruby/bioruby

論文が出ました

Naohisa Goto, Pjotr Prins, Mitsuteru Nakao, Raoul Bonnal, Jan Aerts, Toshiaki Katayama. (2010) BioRuby: bioinformatics software for the Ruby programming language. Bioinformatics 26(20): 2617-2619.

http://bioinformatics.oxfordjournals.org/content/26/20/2617.abstract

（購読者以外でも全文を読めます）

他言語による先行プロジェクト

Perl	BioPerl
Java	BioJava
Python	Biopython

言語により得意分野が異なるので共存

- Open Bioinformatics Foundation (OBF)
  - 情報交換や開発協力、標準化など
  - http://www.open-bio.org/

BioRubyの歴史

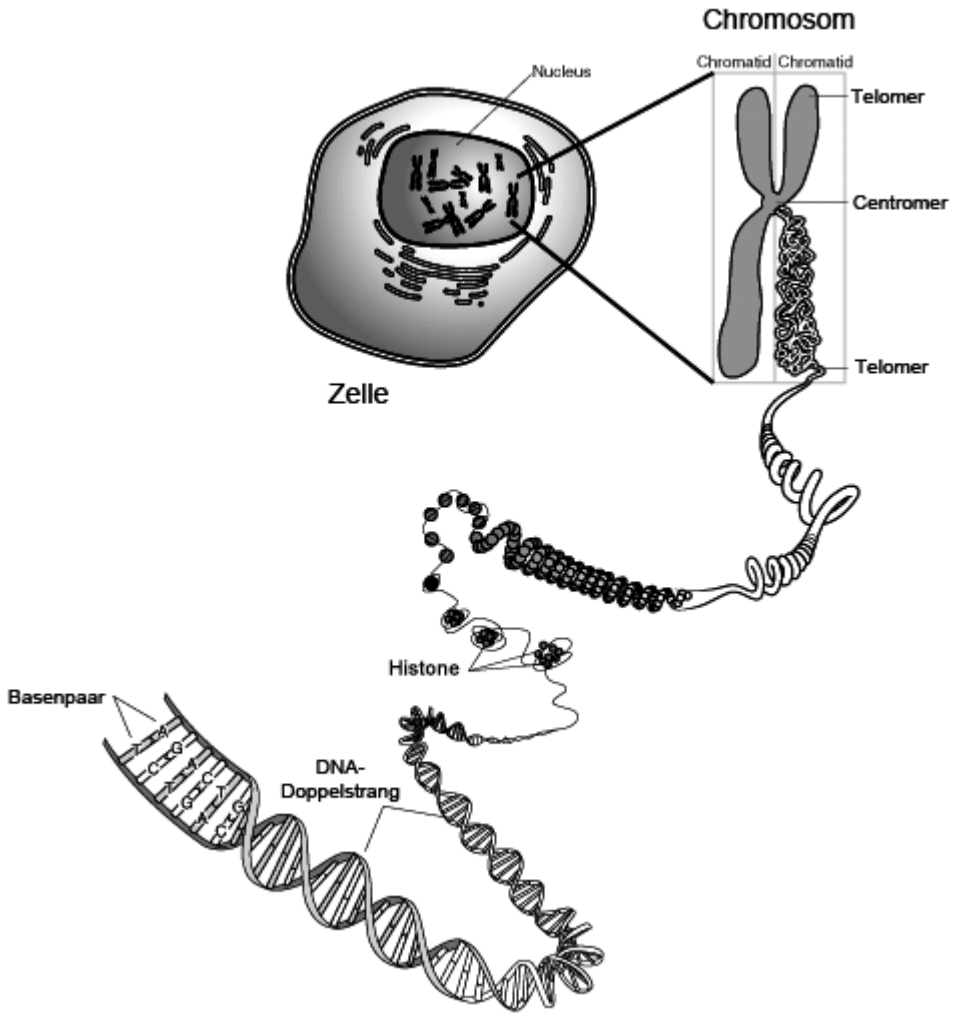
2000/11/21	BioRubyプロジェクト開始（片山、中尾、奥地)
2001/06/21	バージョン0.1リリース
2001/07/19	Bioinformatics Open Source Conference（デンマーク）ライトニングトーク(おくじ)
2001/10/24	バージョン0.3 リリース・CVSレポジトリ開始
2001/11/17	第1回BioRuby宴会（京都）
2001/12/15	バージョン0.3.3 リリース（現存するChangeLogの最初の日付）
2002/02	BioHackathon（南アフリカ）参加（片山）
…	
2005/06	IPA未踏ソフトウェアプロジェクト（2月末まで）
…	
2005/07/09	第4回関西Ruby勉強会にて発表（後藤）
2006/02/24	バージョン1.0リリース・未踏成果報告会＠品川
…	
2006/12	Phyloinformatics Hackathon（アメリカ）参加（片山、後藤）
2007/12/15	バージョン1.2.0リリース・第21回関西Ruby勉強会にて発表（後藤）
2008/2	DBCLS BioHackathon (東京)
2009/3	DBCLS BioHackathon2009 (東京・沖縄)
2009/12/29	バージョン1.4.0 リリース
2010/2	DBCLS BioHackathon2010 (東京)

（敬称略）

BioRubyの規模

- ファイル数
  - ライブラリ本体: 約 230
  - ユニットテスト: 約 120
  - サンプルスクリプト: 約 70
- 行数(コメント・空行を除く)
  - ライブラリ本体: 約 3万5千行
  - ユニットテスト: 約 2万2千行
- class/moduleの数
  - 約 580 クラス・モジュール
  - 約 2800 メソッド(private除く)

生物の持つ情報(1)ゲノム



4種類の塩基

A (アデニン)

T (チミン)

G (グアニン)

C (シトシン)

ひも状

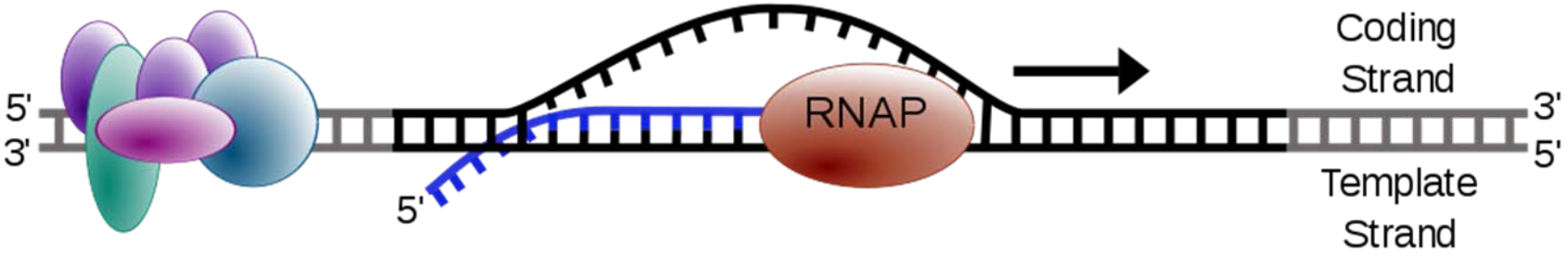
向きがある(5’→3’)

文字列として扱える

図の出典: WikiPedia: 染色体

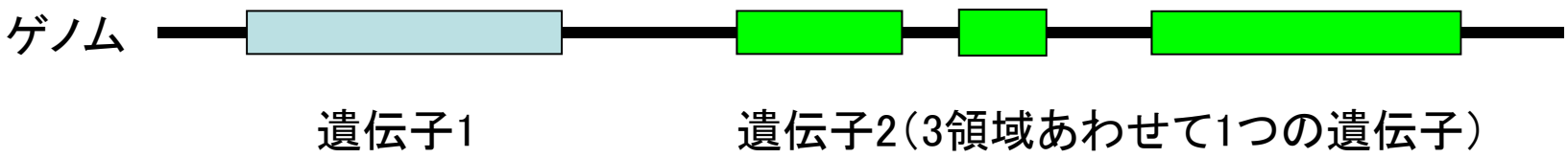
生物の持つ情報(2) 転写・発現

- ゲノムから遺伝子の領域がRNAに「転写」される
  - 遺伝子の「発現」



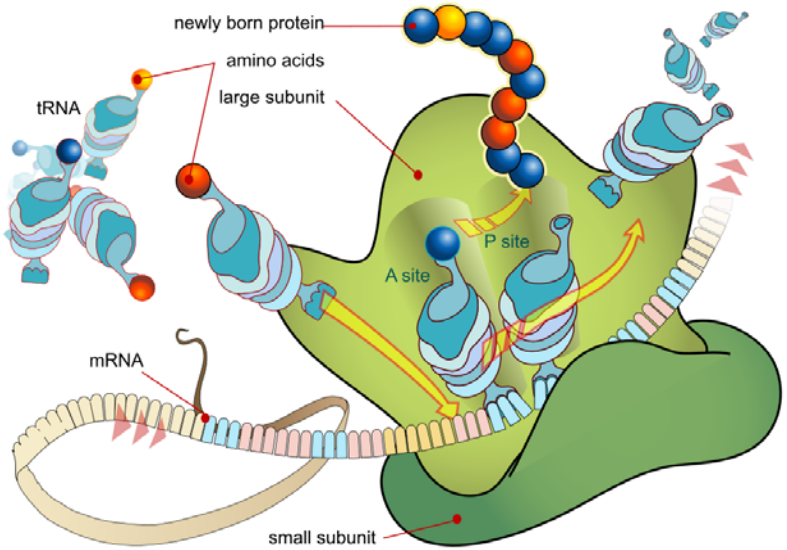
図の出典: WikiPedia: Gene_expression

- ゲノム上の遺伝子領域
  - ある程度の規則性は発見されているが…
  - 発現を制御する領域もゲノム上に存在



生物の持つ情報(3) 翻訳・タンパク質

- mRNAがリボソームでタンパク質に「翻訳」される
- RNA3塩基がアミノ酸20種類に対応
  - 3塩基のことを「コドン」という
  - 翻訳の終了を指示する「終止コドン」もある
- アミノ酸がペプチド結合したものがタンパク質
  - タンパク質も文字列扱いOK



塩基配列データベース

- 世界3か所で管理
  - アメリカ：GenBank　http://www.ncbi.nlm.nih.gov/Genbank/
    （National Center for Biotechnology Informationが運営）
  - ヨーロッパ：EMBL　http://www.ebi.ac.uk/embl/
    （European Bioinformatics Instituteが運営）
  - 日本：DDBJ　http://www.ddbj.nig.ac.jp/
    （国立遺伝学研究所が運営）
- データは3か所で常に相互に交換
  - IDは3か所共通管理＝どれか1か所に登録すればOK
- 新規塩基配列は登録がほぼ必須
  - すべての学術雑誌が、塩基配列を登録してそのID（アクセッション番号）を論文に掲載することを求めている
- 無償公開、データ利用の制限がほとんどない

データの例（GenBank）

テキスト形式、1エントリ1配列

配列だけでなく付加情報も付いてくる

```
LOCUS       HUMADH1CB               1400 bp    mRNA    linear   PRI 08-JUN-1995
DEFINITION  Homo sapiens class I alcohol dehydrogenase (ADH1) alpha subunit
            mRNA, complete cds.
ACCESSION   M12271
VERSION     M12271.1  GI:178091
KEYWORDS    ADH1 gene; alcohol dehydrogenase; alcohol dehydrogenase I;
            dehydrogenase.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Catarrhini;
            Hominidae; Homo.
REFERENCE   1  (bases 1 to 1400)
  AUTHORS   Ikuta,T., Szeto,S. and Yoshida,A.
  TITLE     Three human alcohol dehydrogenase subunits: cDNA structure and
            molecular and evolutionary divergence
  JOURNAL   Proc. Natl. Acad. Sci. U.S.A. 83 (3), 634-638 (1986)
   PUBMED   2935875
COMMENT     Original source text: Homo sapiens (clone: pUCADH-alpha-15L) liver
            cDNA to mRNA.
            A draft entry and printed copy of the sequence in [1] were kindly
            provided by A.Yoshida, 30-MAY-1986.
            The other human class I ADH1 alpha subunit sequence is found under
            accession M11307.1
```

GenBank形式(続き)

```
FEATURES             Location/Qualifiers
     source          1..1400
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
                     /map="4q21-q23"
                     /clone="pUCADH-alpha-15L"
                     /tissue_type="liver"
     gene            1..1400
                     /gene="ADH1"
     mRNA            <1..1400
                     /gene="ADH1"
                     /note="G00-119-650"
     CDS             16..1143
                     /gene="ADH1"
                     /EC_number="1.1.1.1"
                     /note="alpha subunit"
                     /codon_start=1
                     /product="alcohol dehydrogenase 1"
                     /protein_id="AAA68131.1"
                     /db_xref="GI:178092"
                     /db_xref="GDB:G00-119-650"
                     /translation="MSTAGKVIKCKAAVLWELKKPFSIEEVEVAPPKAHEVRIKMVAV
                     GICGTDDHVVSGTMVTPLPVILGHEAAGIVESVGEGVTTVKPGDKVIPLAIPQCGKCR
                     ICKNPESNYCLKNDVSNPQGTLQDGTSRFTCRRKPIHHFLGISTFSQYTVVDENAVAK
                     IDAASPLEKVCLIGCGFSTGYGSAVNVAKVTPGSTCAVFGLGGVGLSAIMGCKAAGAA
                     RIIAVDINKDKFAKAKELGATECINPQDYKKPIQEVLKEMTDGGVDFSFEVIGRLDTM
                     MASLLCCHEACGTSVIVGVPPDSQNLSMNPMLLLTGRTWKGAILGGFKSKECVPKLVA
                     DFMAKKFSLDALITHVLPFEKINEGFDLLHSGKSIRTILMF"
```

GenBank形式(続き)

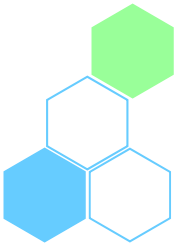
```
ORIGIN      52 bp upstream of PvuII site; chromosome 4q21.
        1 gaagacagaa tcaacatgag cacagcagga aaagtaatca aatgcaaagc agctgtgcta
       61 tgggagttaa agaaaccctt ttccattgag gaggtggagg ttgcacctcc taaggcccat
      121 gaagttcgta ttaagatggt ggctgtagga atctgtggca cagatgacca cgtggttagt
      181 ggtaccatgg tgaccccact tcctgtgatt ttaggccatg aggcagccgg catcgtggag
      241 agtgttggag aaggggtgac tacagtcaaa ccaggtgata aagtcatccc actcgctatt
      301 cctcagtgtg gaaaatgcag aatttgtaaa aacccggaga gcaactactg cttgaaaaac
      361 gatgtaagca atcctcaggg gaccctgcag gatggcacca gcaggttcac ctgcaggagg
      421 aagcccatcc accacttcct tggcatcagc accttctcac agtacacagt ggtggatgaa
      481 aatgcagtag ccaaaattga tgcagcctcg cctctagaga aagtctgtct cattggctgt
      541 ggattttcaa ctggttatgg gtctgcagtc aatgttgcca aggtcacccc aggctctacc
      601 tgtgctgtgt ttggcctggg aggggtcggc ctatctgcta ttatgggctg taaagcagct
      661 ggggcagcca gaatcattgc ggtggacatc aacaaggaca aatttgcaaa ggccaaagag
      721 ttgggggcca ctgaatgcat caaccctcaa gactacaaga aacccatcca ggaggtgcta
      781 aaggaaatga ctgatggagg tgtggatttt tcatttgaag tcatcggtcg gcttgacacc
      841 atgatggctt ccctgttatg ttgtcatgag gcatgtggca caagtgtcat cgtaggggta
      901 cctcctgatt cccaaaacct ctcaatgaac cctatgctgc tactgactgg acgtacctgg
      961 aagggagcta ttcttggtgg ctttaaaagt aaagaatgtg tcccaaaact tgtggctgat
     1021 tttatggcta agaagttttc attggatgca ttaataaccc atgttttacc ttttgaaaaa
     1081 ataaatgaag gatttgacct gcttcactct gggaaaagta tccgtaccat tctgatgttt
     1141 tgagacaata cagatgtttt cccttgtggc agtcttcagc ctcctctacc ctacatgatc
     1201 tggagcaaca gctgggaaat atcattaatt ctgctcatca cagattttat caataaatta
     1261 catttggggg ctttccaaag aaatggaaat tgatgtaaaa ttatttttca agcaaatgtt
     1321 taaaatccaa atgagaacta aataaagtgt tgaacatcag ctggggaatt gaagccaata
     1381 aaccttcctt cttaaccatt
//
```

Fasta形式

配列データだけを扱う場合のシンプルな形式
>から始まる行に配列のIDや説明など
その直後に配列データ(配列データ中の改行は無視される)
もちろんBioRubyでは入出力ともにサポート

```
>M12271 human ADH1 alpha subunit mRNA
gaagacagaatcaacatgagcacagcaggaaaagtaatcaaatgcaaagcagctgtgctatgggagttaa
agaaaccctttccattgaggaggtggaggttgcacctcctaaggcccatgaagttcgtattaagatggt
ggctgtaggaatctgtggcacagatgaccacgtggttagtggtaccatggtgaccccacttcctgtgatt
ttaggccatgaggcagccggcatcgtggagagtgttggagaaggggtgactacagtcaaaccaggtgata
aagtcatcccactcgctattcctcagtgtggaaaatgcagaatttgtaaaaacccggagagcaactactg
cttgaaaaacgatgtaagcaatcctcaggggaccctgcaggatggcaccagcaggttcacctgcaggagg
aagcccatccaccacttccttggcatcagcaccttctcacagtacacagtggtggatgaaaatgcagtag
ccaaaattgatgcagcctcgcctctagagaaagtctgtctcattggctgtggattttcaactggttatgg
gtctgcagtcaatgttgccaaggtcaccccaggctctacctgtgctgtgtttggcctgggaggggtcggc
ctatctgctattatgggctgtaaagcagctggggcagccagaatcattgcggtggacatcaacaaggaca
aatttgcaaaggccaaagagttgggggccactgaatgcatcaaccctcaagactacaagaaacccatcca
ggaggtgctaaaggaaatgactgatggaggtgtggatttttcatttgaagtcatcggtcggcttgacacc
atgatggcttccctgttatgttgtcatgaggcatgtggcacaagtgtcatcgtaggggtacctcctgatt
cccaaaacctctcaatgaaccctatgctgctactgactggacgtacctggaagggagctattcttggtgg
ctttaaaagtaaagaatgtgtcccaaaacttgtggctgattttatggctaagaagttttcattggatgca
ttaataacccatgttttacctttgaaaaaataaatgaaggatttgacctgcttcactctgggaaaagta
tccgtaccattctgatgttttgagacaatacagatgttttcccttgtggcagtcttcagcctcctctacc
ctacatgatctggagcaacagctgggaaatatcattaattctgctcatcacagattttatcaataaatta
catttgggggctttccaaagaaatggaaattgatgtaaaattatttttcaagcaaatgtttaaaatccaa
atgagaactaaataaagtgttgaacatcagctggggaattgaagccaataaaccttccttcttaaccatt
```

BioRubyのファイル形式自動認識



- Bio::FlatFile でファイル形式を自動認識
  - 使用例:

```
Bio::FlatFile.open("file.gbk") do |ff|
    ff.each { |e|  p e.naseq }
end
```

- 利点
  - ファイル形式をいちいち指定しなくて済む
  - ダックタイピングを活用できる
    - 同じようなデータ形式なら同じプログラムで済ませられる
- 欠点
  - 多少オーバーヘッドがある
    - 多量の大きいファイルを読む場合に影響

DNAシーケンサー

- DNAを「読む」実験装置
- 従来型（サンガー法）
  - Applied Biosystems社(略称ABI)の製品
  - 1サンプルあたり最長700-1000塩基読める
    - 長い配列は切断して複数に分けて読み、後で情報処理してつなぐ
  - 最大約9,000本/日/1台
    - （装置により性能は異なる）

次世代シーケンサー

- イルミナ Illumina Genome Analyzer
  - 1本あたり最長100-150塩基
  - 最大6～20億本/1ラン(1週間～10日)
  - つまり1回あたり6GB～20GB読める
  - コストは100万円～
- ロシュ 454 GS Titunium
  - 1本あたり最長400-600塩基
    - ただし精度があまり良くない
  - 最大100万本/1ラン(10時間)
  - つまり1回あたり400MB～600MB読める
  - コストは100万円～

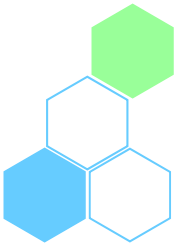
次世代シーケンサー対応に向けて

- 大量データへの対応
- 各種データフォーマットへの対応
- 既に多数出ている次世代シーケンサー向けアプリケーションへの対応

今後の予定

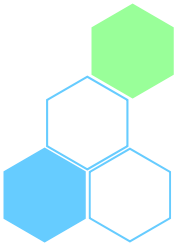
- プラグイン化
  - ユーザーが後からモジュールを追加
  - 開発の敷居を下げる
- ドキュメントの充実
  - HOWTO, チュートリアルetc.

Google Summer of Code とは？



- Googleのオープンソース支援のひとつ
- 学生に資金提供してコードを書いてもらう
  - 5000ドルを2回に分けて提供
  - 別途500ドルはプロジェクトに提供
- 学生はオープンソースプロジェクトに参加
  - コードを書くのが「仕事」
- コーディング期間はアメリカの夏休み
  - 2010年は5/24-8/16
- http://code.google.com/soc/

私の体験



- 2009年 NESCent (National Evolutionary Synthesis Center, アメリカの国立研究機関)
  - サブのメンターとして参加
  - コードのレビューや質問への回答など
  - 採点・評価にはほとんど関与せず
  - GoogleのTシャツは貰えた
- 2010年 OBF (Open Bioinformatics Foundation)
  - メンターとして参加
  - 採点・評価にも関与
  - その他もろもろ

用語集

- メンター (mentor)
  - 学生を指導し助言を与える人
  - 学生1人につき1人以上
    - 1人のメンターが複数学生を指導することは可能
  - オープンソースプロジェクトに所属
    - 複数プロジェクト掛け持ちもOK
- メンター組織 (mentoring organization)
  - Google Summer of Code に参加しているオープンソースのプロジェクトのこと
  - 各メンターはここに属する形式

スケジュール (2010年)

- 2月: 開催アナウンス
- 3/8-12 メンター組織募集
- 3/18 メンター組織決定
  - メンター組織は学生応募用のプロジェクトアイデアを掲載
- 3/29-4/9 学生応募期間
- 4/21 学生の提案書(proposal)の採点締切
  - 複数組織に応募している場合は代表者のIRC会議で決定
  - 学生ごとにメンター割り当て
- 4/26 採用者発表
- 4/26-5/9 “Community Bonding Period”
  - コミュニティに慣れ親しむ
  - ソースやドキュメントを読むなどの準備

スケジュール (2010年)

- 5/24 コーディング期間スタート
- 7/12-16 中間評価(学生・メンターの両方が行う)
- 8/9「筆を置く」のを推奨する日("pencil-down date")
  - この日までにコーディングを終了することを推奨
  - 残り1週間はドキュメントやテストを仕上げる
- 8/16 最終日
- 8/16-20 最終評価(学生・メンターの両方が行う)
- 8/23 Googleから結果発表
- 8/30- Googleに書いたコードを提出
- 10/23-24 Mentor Summit at Google

学生の採択基準

- 学生の評価・採点は各組織に任されている
- Googleは採点に口出ししない
  - ただし評価をちゃんとやっているかは見ているはず
    - いい加減な組織は来年採用されないかも？
- Googleは各組織毎に人数を割り当てる
  - 割当人数は組織毎の応募総数を参考にしているらしい
  - 当初（採点締切の数日前）概数割り当て、後に若干の追加
  - OBFの場合、当初5名、後に追加1名

OBFでのプロポーザル採点基準

- プロポーザルの内容
  - やるべきことを理解できているか
- コーディング経験・能力
- コミュニケーションが取れそうか？
  - 場合によっては電話インタビューあり
- プロジェクトに専念できるか？
  - 他の仕事や授業などの有無
    - 日本の学生はたいへん不利

参加するには(1)学生として

- あらかじめプロジェクトに目星をつけておく
  - 事前にメーリングリストに投稿するなどすると効果大
- 3月末～4月初、プロジェクトを選んで提案書を書く
  - 締切ぎりぎりの応募は避けよう
  - 英語で書く必要あり
  - 複数応募は可能
- 採用されたら、コードを書く
  - 多くの場合、週1回程度進捗を報告する必要がある
  - あとはただひたすらコードを書く

参加するには(2)プロジェクト・メンターとして

- プロジェクトのプロポーザルを書いてGoogleに応募
  - 当然ながら英語で書く必要あり
- 学生応募用のプロジェクトアイデアを出す
  - 学生にコードしてもらいたいことを出す
  - たくさん出すと学生の目に留まることが多いかも？
- メンターに登録
- 学生の評価・採点に加わる