Technical report on variant callers

Deep learning in bioinformatics

https://github.com/gokceneraslan/awesome-deepbio

Same review in Bioinformatics (Oxford Press).

Overall deep learning in bioinformatics and computational biology is growing exponentially since the early 2010's, this means that there are not a lot of papers on each problem (about 3 or 4 in sequencing).

Definitions

This section contains the definitions of technical terms used in this report, please feel free to refer back to them as you go through the report.

Genetics

- *Ploidy* : number of sets of chromosomes in a cell.
- *Diploid* : organism with two sets of chromosomes, thus having two alleles that can be the same or different.
- *Haploid* : other name for monoploid, only a single set of chromosomes in a cell.
- *Homozygote* : the two alleles are the same.
- *Heterozygote* : the two alleles are different.

Types of mutations

- *Single Nucleotide Polymorphism (SNP)* : change in the DNA of a nucleotide for another one (e.g. replacing a C with a G at a specific position). SNPs also have the constraint to be present in more than 1% of the population.
- *Single Nucleotide Variation (SNV)* : same as SNP, without the frequency constraint.
- *Base pair (bp)*: unit of measurement in the genome, counts the number of nucleotides (i.e. a deletion of "ACGT" would be a 4bp deletion).
- *Small variations* : variation of a few nucleotides in the DNA, either insertions/deletions of less than 10bp (this limit is blury) or SNPs.
- *Copy Number Variation (CNV)* : some portions of the DNA are repeated inside the DNA, the number of time they are repeated has a strong influence on the subject.
- *Structural Variation (SV)* : large variations in the DNA, they are either inversions, translocations, copy number variations, large insertions/deletions or duplications.

- *Germline mutation* : Mutation that is passed down to the next generation (such as eye color).
- *Somatic mutation* : Mutation that is *not* passed down to the next generation (such as cancer).

Tools and techniques

- *Variant caller* : Tool used to detect variations in the DNA, either small variations or structural variations.
- *Population genomics* : Use of data from other samples from the same population in order to have priors in the analysis of a sample.

Sequencing

- *Read* : sequencer do not work by sequencing the a whole DNA molecule at once, but splits it in small sequences called reads.
- *Nx coverage* : Nx coverage means that each part of the genome is expected to have about N reads containing it. The coverage thus has a linear relationship with the price of sequencing.
- *reference genome* : sequenced genome used as a ground truth, not necessarily coming from a single individual.
- *Alignment* : mapping reads to a reference genome is called alignment.
- *SAM format* : Sequence Alignment Map, file format containing information about aligned reads in ASCII.
- *BAM format* : Binary Alignment Map, a compressed SAM file.
- *de novo sequencing* : sequencing of a genome without using a reference genome.
- *Variant Call Format (VCF)* : file format containing information about the variants discovered.

Small Variation Callers

Problem statement

Detecting small mutations is still an open problem.

The current method of measurement is precision using a fixed recall, i.e. aiming for 99 precision with a 90% recall.

The input data is usually aligned reads in BAM format. This can be done because since variations are small by definition they are thus not too far away from a reference genome and can be properly aligned.

In SNP discovery competitions the input is usualy FASTA/FASTQ files and the alignment is left to be done by competitors (usually using Bowtie or BWA).

The output of this task is a VCF file, with the location and types of the variants (indel or SNP).

Current limitations

The current methods requires a lot of fine tuning by experts for each sample, depending on the sequencing technology used, the physical sequencing machine used (two Illumina sequencers of the same model will still require different tunings in variant calling) and the species of the sequenced sample.

This leads to think that the methods are not robust and do not discover invariants in the data.

This also means that getting rid of this tuning process (or at least doing it only once) would be a great innovation.

Current models also make assumptions that are known to be false about the data, in particular sequencing errors are assumed to be independent in the model when we know that they are in fact correlated.

Available datasets

Genome In A Bottle consortium: contains 5 genomes with deep (300x and 100x) coverage sequenced with 12 different technologies, current standard in the domain.
This dataset is used by the FDA precision medicine challenge.

1000 genome project: 1000 genomes with low (4x) coverage with detailed annotations, spearheaded by the Broad Inisitute.

Current tools and methods

Most of the current methods use simple statistical models (Naive Bayes + Gaussian Mixtures), the current gold standard is GATK's HaplotypeCaller described in GATK's best practices.

DeepVariant is a current challenger that scored the best SNP F-score in the FDA precision challenge (99.9587%) despite performing lower than the other tools in indels.

Here are the results of the top scoring teams (ranked by SNP-Fscore) :

Label	Submitter	Organization	SNP-Fscore	SNP-recall	SNP-precision	INDEL-Fscore	INDEL-recall	INDEL-precision
rpoplin-dv42	Ryan Poplin et al.	Verily Life Sciences	99.9587	99.9447	99.9728	98.9802	98.7882	99.1728
hfeng-pmm3	Hanying Feng et al.	Sentieon	99.9548	99.9339	99.9756	99.3628	99.0161	99.7120
hfeng-pmm1	Hanying Feng et al.	Sentieon	99.9496	99.9227	99.9766	99.3397	99.0289	99.6526
dgrover-gatk	Deepak Grover	Sanofi-Genzyme	99.9456	99.9631	99.9282	99.4009	99.3458	99.4561
hfeng-pmm2	Hanying Feng et al.	Sentieon	99.9416	99.9254	99.9579	99.3119	99.0152	99.6103
jli-custom	Jian Li et al.	Roche	99.9382	99.9603	99.9160	99.3675	99.0788	99.6580
bgallagher-sentieon	Brendan Gallagher et al.	Sentieon	99.9296	99.9673	99.8919	99.2678	99.2143	99.3213
raldana-dualsentieon	Rafael Aldana et al.	Sentieon	99.9260	99.9131	99.9389	99.1095	98.7566	99.4648

We can see that dgrover-gatk is the best at INDEL F-score and recall while having ‘high scores’ in all other categories.
GATK thus seems to keep its place as the de facto gold standard.

Nevertheless, since many pipelines are using GATK it seems that the art here is to properly parametrize GATK and to have properly preprocessed the data.

Since DeepVariant seems not to have invested a lot in preprocessing (to their own admission) and that there focus was on SNP Fscore it stands to reason that there is room for improvements

More details about the results can be found on the FDA precision webpage.

Use of Deep Learning

One of the main hurdles in deep learning is finding a representation of the input data that is suitable for deep learning algorithms (preprocessing), once a representation has been chosen the type of neural network to use comes easily.

I believe that little work in SNP discovery was done using deep learning because of a lack of such a representation.

DeepNANO

DeepVariant

The team lead by Ryan Poplin found such a representation which allowed them to use a convolutional neural network architecture (Inception V2, which is no longer the state of the art in convolutional neural networks, Densenet and XNet currently are as of 2017/09).

DeepVariant, leverages Inception, a neural network trained for image classification by Google Brain, by encoding reads around a candidate SNP as a 221x100 bitmap image, where each column is a nucleotide and each row is a read from the sample library [257]. The top 5 rows represent the reference, and the bottom 95 rows represent randomly sampled reads that overlap the candidate variant. Each RGBA (red/green/blue/alpha) image pixel encodes the base (A, C, G, T) as a different red value, quality score as a green value, strand as a blue value, and variation from the reference as the alpha value.

This method has many advantages:

- Takes into account correlation in sequencing errors.
- Is scalable from machine to machine.
- Is scalable from sequencing technology to sequencing technology.
- Is scalable from species to species.
- Achieved best truth score at the 2016 FDA precision challenge.

This method is nevertheless improvable in two ways:

- The representation was an image where data was encoded in the pixel values, which is not the closest fit to the data. I believe this was done in order to speed up the process as Google has a lot of knowledge in training CNN for image tasks. Using a representation with higher dimension by encoding the inputs as binary classes for each nucleotide would make more sense, this approach has been proposed by the original paper and confirmed in email exchanges with them.
- Using a better convnet architecture could bring small improvements, even though I don't belive this would bring a lot value, as the main task here is in the representation of the data.

DeepVariant is expected to be open sourced in the near future according to the Google Cloud Engine webpage.

Doing the proposed improvements has been tried by Jason Chin and is explained in details in "Simple Convolutional Neural Network for Genomic Variant Calling with TensorFlow". I still believe that his implementation can be improved upon.

Campagne's lab

I am mentionning these papers because they were in the "Opportunities. . . " and come up in google searches for "SNP calling" and "Deep Learning".

To my opinion they did not bring any fundamental ideas relative to the other articles, their main contributions are an opensource deep learning implementation (which should be beaten by DeepVariant) and the use of data augmentation which is already a standard practice in Machine Learning.

Adaptive Somatic Mutations Calls with Deep Learning and Semi-Simulated Data Remi Torracinta, Laurent Mesnard, Susan Levine, Rita Shaknovich, Maureen Hanson, Fabien Campagne bioRxiv 079087; doi: https://doi.org/10.1101/079087

Training Genotype Callers with Neural Networks Rémi Torracinta, Fabien Campagne bioRxiv 097469; doi: https://doi.org/10.1101/097469

Conclusion

I belive that deep learning would be an appropriate method here and that there ae still improvements to be done on the work done by Poplin et al..

This could be done by December in the worst case and could be worthy of publication in either Bioinformatics or a workshop at ICLR.

The advantages would be to :

- Build a working genomic pipeline that could later be reused for Structural variant discovery.
- Get hands on experience with the data at a low cost.
- Be doable quickly because all data is currently freely accessible.
- Get a publishable result within 3 months of the beginning of the PhD.
- Create a working relationship with Ryan's team (headed by Mark DePristo which is head of genomics at Google Brain and Verily).
- Gather information on Structural variant discovery.

Structural Variant Callers

This problem is much harder and reap for deep learning.

Data is as heterogenous as is possible and will be a challenge to represent in a Deep Learning friendly way.

Reviews

Lin K., Smit S., Bonnema G., Sanchez-Perez G., de Ridder D. (2014). Making the difference: integrating structural variation detection tools. Brief. Bioinform. 10.1093/bib/bbu047

Great review of structural variations

Tattini L, D'Aurizio R, Magi A. Detection of Genomic Structural Variants from Next-Generation Sequencing Data. Frontiers in Bioengineering and Biotechnology. 2015;3:92. doi:10.3389/fbioe.2015.00092.

Other review of SV callers. Broader and with more explanations, but does not go into ensemble models.

Tools

Chen, X. et al. (2016) Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics, 32, 1220-1222. doi:10.1093/bioinformatics/btv710

Manta overview, just presents the results and says that they are cool. Uses contig graph, then some stats on that data (to be developped).

Ye K. et al. . (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics , 25, 2865–2871.

Rausch T. et al. . (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics , 28, i333–i339.

Optional: Deep learning in genomics publications

Some review has also been done in Bioinformatics (Oxford press) about the kind of deep learning papers accepted.

I selected the Deep Learning papers of the last two issues and it appears to me that very little time is spend on the architecture, most of the work is either in the preprocessing or even just applying a neural network without any preprocessing.