Quantum genetic medicine

I. QUANTUM GENETIC MEDICINE

Until now, medicine has largely relied on one-size-fits-all diagnosis and treatment options. Needless to say, everyone's different, manifest in their unique genotype. However, with access to individuals' unique genetic makeup, the next generation of medicine will become highly personalised, catering for individual genetic differences. People with different genetic predispositions, mutations, or traits will be able to have treatment options tailored to them. Different cancer types, which are genetically distinct from one another, may be genetically targeted.

With the ability to sequence individuals' genomes extremely cheaply, we will open up entirely new medical possibilities for genetically-personalised treatments – the era of *genetic medicine*, the next revolution in medicine.

Doing so will require highly complex processing of massive amounts of genetic and drug information, requiring demanding computational resources. We forsee quantum computing as playing a central role in the processing pipeline of genetic drug development, some initial ideas for which we present here.

A. The human genome project

The first complete mapping of the human genome was a major scientific achievement [**?**], opening up entirely new avenues for medical research that were previously never possible. It was, however, an extraordinarily expensive undertaking, costing on the order of $1b.

Since then, genetic sequencing tools have undergone a massive technological transformation and are now available as commodity hardware at price-points accessible to any well-resourced bioscience lab. This now enables sequencing the human genome orders of magnitude cheaper than the first attempt. It is now possible to map the genome for ∼$1,000's.

Following its own technological Moore's Law, one can reasonably anticipate this process becoming sufficiently cheap that in the near future it will become economically viable for every individual to have their own personal genome fully sequenced and available for medical use.

1. *Short-read sequencing*

The major technological transition that has enabled this rapid progress is the adoption of next-generation sequencers (NGS) based on *short-read* technology. Using this process, a DNA sequence is not mapped exhaustively from beginning to end, but rather is chemically deconstructed into an enormous number of *short-reads* – small genetic segments, each on the order of ∼50 base-pairs in length. Having prepared an enormous pool of such short-reads, they are then sequenced in parallel, yielding a large database of N short strings, corresponding to small segments of the larger genome. The task then is to reconstruct a complete genome from this data.

To achieve this, there are two approaches that are commonly employed – *de novo* assembly, and *mapping*.

2. *De novo sequencing*

In *de novo* sequencing we treat the short-read data as a jigsaw puzzle that must be reassembled. We define an overlap threshold, n, and upon comparing every string against every other, look at whether their ends overlap consistently by at least n base-pairs. When a match is found, the respective short-reads are merged into a larger *contig*. This process continues until no further sufficiently-overlapping strings are found. At this point we should, at least in principle, have a fully reassembled genome, or at least very large contigs belonging to it (assuming a sufficiently large pool of short-reads to begin with).

Open-source software for implementing *de novo* assembly of short-read data is available. The well-known *Velvet* package [**?**] does this graph theoretically using de Bruijn graph representations for contigs/reads, with very efficient computational resource requirements.

The number of comparisons between short-reads scales only as $O(N^2)$, and employing hash-table representations for short-reads, the lookups for each individual comparison can be efficiently implemented in $O(1)$ time. The confidence that two contigs/reads actually overlap increases exponentially with n, but with increasing n comes a reduction in the number of matches that will be identified – the tradeoff between *sensitivity* and *specificity* in contig reconstruction.

The *de novo* approach is extremely powerful, as it does not require any genomic reference. Rather, we can reconstruct a genome *ab initio*, with sufficient read data.

3. *Mapping*

The failing of *de novo* sequencing is that the pool of short-read data must be sufficiently large that all the pieces in the jigsaw puzzle are present, such that there are no gaps between neighbouring pieces – contig 'islands' will remain isolated from other contigs. However, resources are of course always finite, as too is the number of short-reads available for reconstruction.

The other approach, *mapping*, is to use an existing genome as a reference. Rather than piece short-reads together, we compare them against this reference to infer their relative locations in the genome. When doing so, we allow some error threshold in the matching. Specifically, we require the Hamming distance between a short-read and an equally-long segment of the reference be below some threshold. The flexiblity offered by this threshold acommodates for mutational differences between the reference and the short-reads, which is ultimately what we wish to characterise.

To illustrate the power of this approach, consider the following. Only a miniscule fraction of the genome differs from one individual to the next – almost all of it is identical. Thus, mapping allows us to identify the mutational differences between individuals. Importantly, unlike *de novo* sequencing, short-reads can be mapped to the reference even with incomplete data – contig 'islands' needn't remain so. Non-overlapping short-reads can still be successfully mapped, making this approach applicable (but potentially incomplete) even with datasets insufficiently large for *de novo* sequencing to be effective.

In the case of cancer diagnosis, for example, a cancer cell's genetic makeup is virtually identical to that of its host, modulo some small number of mutations that characterise the cancer. Using the mapping approach we can quickly identify these mutations, hence understanding the genetics of the underlying cancer, potentially opening opportunities for genetically-targeted treatment.

As with *de novo* sequencing, efficient open-source software for efficiently implementing mapping using commodity hardware is available [**?**].

B. Genetic medicine

At its simplest, the design of genetically-tailored medical treatments will require understanding the interactions between drug compounds and genetic processes at the molecular level. For example, we might desire that a drug modifies gene expression for specific genes, or the transcription of DNA to specific proteins. In the case of cancer treatment, we may wish to inhibit the function of biochemical processes reliant on genes exhibiting particular cancerous mutations, whilst not affecting healthy, unmutated ones.

The drug design process for conventional medicines is an incredibly tedious one, requiring massive latitudinal and longitudinal studies of the effects of drugs on physiological symptoms. However, in the case of individually tailored medicine such studies are clearly not possible. This will require replacing human studies with accurate biomolecular-level simulations of these chemical processes. At this scale, interactions are inevitably quantum mechanical in nature, which must be accomodated for in simulations.

The key goal is to properly simulate the interaction of candidate drug compounds, ligands, or functional groups with biochemical processes at the molecular level. Classical techniques for such simulations include estimating electron densities using density functional theory (the potential energy surface) of compounds, from which useful properties such as bonding affinities may be determined.

C. Quantum chemistry

Despite the existence of classical estimation techniques, molecular-level interactions are necessarily quantum mechanical in nature. For this reason, classical simulation techniques are limited to approximations that ignore realistic and important quantum effects, since this would require exponential classical computational resources in general (i.e in general, simulating quantum systems is **BQP**-complete).

Using standard quantum chemistry techniques, one can construct accurate Hamiltonians describing the evolution of molecular systems at the quantum level. Efficiently implementing such simulations can then be performed on quantum computers using standard, efficient Hamiltonian simulation quantum algorithms (Sec. **??**) for simulating the Schrödinger equation,

$$\hat{H}|\psi(t)\rangle = i\hbar\frac{\partial}{\partial t}|\psi(t)\rangle, \tag{1}$$

where $\hat{H}$ is the system's Hamiltonian, describing the time-dependent evolution of the state as,

$$|\psi(t)\rangle = e^{-\frac{i\hat{H}t}{\hbar}}|\psi(0)\rangle. \tag{2}$$

These Hamiltonians can be constructed *ab initio*, combining relevant one- and two-body interactions, such as Coulomb potential, kinetic energy, spin- and spin-orbit coupling, and magnetic and electric dipole terms appropriately. A net interaction Hamiltonian may be obtained by summing over the different one- (i) and two-body (j,k) terms for all particles comprising the system,

$$\hat{H}_{\text{int}} = \sum_i \hat{H}_i + \sum_{j>k} \hat{H}_{j,k}. \tag{3}$$

The dominant one-body interactions are kinetic energy terms, of the form,

$$\begin{aligned}\hat{H}_i^{(\text{kinetic})} &= \frac{\hat{p}^2}{2m}\\ &= -\frac{\hbar^2}{2m_i}\nabla^2_{\vec{r}_i},\end{aligned} \tag{4}$$

where $\hat{p}$ is the momentum operator, m_i is mass, $\vec{r}_i$ is the position vector, and,

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}. \tag{5}$$

The dominant two-body terms are Coulomb (electrostatic) interactions,

$$\hat{H}_{j,k}^{(\text{Coulomb})} = \frac{q_j q_k}{4\pi\epsilon_0|\vec{r}_j - \vec{r}_k|}, \tag{6}$$

where q_j denotes charge.

D. Quantum drug trial simulations

These approaches borrowed from quantum chemistry allow simulation of the interaction between drug compounds and molecular biochemical interactions. The next step is to perform this process against a huge library of candidate drug compounds, ligands or functional groups, from which, we hope, some candidates will exhibit desired characteristics, such as bonding affinities.

To achieve this, let use construct a classical algorithm for exhaustively constructing and enumerating organic drug molecules or functional groups, up to some cutoff size. The length of this list grows exponentially with the cutoff size. Next we implement this algorithm unitarily (which is always possible, writing the classical circuit as a reversible one), and construct it as a quantum oracle such that every input state (which acts as a pointer to an element in the list) yields as output a qubit representation of the corresponding drug compound. Specifically, the oracle implements a transformation of the form,

$$\hat{U}_{\text{oracle}}|i\rangle|0\rangle^{\otimes n} \to |i\rangle|\psi_i\rangle^{(n)}_{\text{molecule}}, \tag{7}$$

where i is the index (pointer) to the drug element in the enumeration, and n is the number of qubits encoding the representation of the drug molecule[1].

Having contructed this oracle, we feed the output qubit molecule description into the Hamiltonian simulation algorithm, with exponential speedup, yielding as output a 'score' characterising some desired property of the interaction, such as bonding affinity. The joint oracle-simulation algorithm is finally embedded within a quantum search subroutine, which searches over the input space enumerating the set of drug candidates for scores above some desired threshold. This yields a quadratic enhancement in the number of drug candidates that can be simulated in parallel.

The complete algorithmic pipeline is shown in Fig. 1.

E. In the cloud

This quantum processing pipeline lends itself well to several important cloud-based protocols. The pipeline comprises several distinct subroutines, which might be outsourced or distributed independently of one another.

An R&D organisation might specialise in the algorithmic generation (or hard-coding) of candidate drug compounds, which they would like to retain as a trade secret, but desire to license to third-party drug designers, who wish to employ them in their computational pipeline. The quantum search and Hamiltonian simulation subroutines might similarly be outsourced or distributed over the cloud to vendors specialising in the implementation of these particular subroutines.

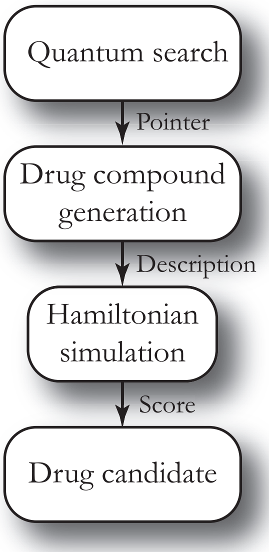

FIG. 1: Pipeline for quantum-enhanced genetic drug development. A quantum search algorithm searches over algorithmically-generated drug compounds, for each performing Hamiltonian simulation, to determine if the compound's score is above a threshold. The quantum search subroutine yields a quadratic enhancement in the number of drugs that can be simulated in parallel, and the Hamiltonian simulation subroutine yields an exponential enhancement over classical simulation techniques.

In an era where the genetic composition of every individual is fully characterised, data security will be of utmost importance – medical confidentiality will be lifted to an entirely new plane when people's genes are at stake, as the nefarious uses and misuses for obtaining other people's genomes are immense. This provides a perfect example for the value of encrypted quantum computation (Sec. ??). If a medical lab is outsourcing some aspects of a computation involving clients' genetics, it is paramount that this be obscured from third-parties performing the computations in the interest of medical confidentiality. Computational efficiency aside, this consideration on its own already justifies implementing such simulation pipelines quantum mechanically, as encrypted classical computation is effectively unviable in general.