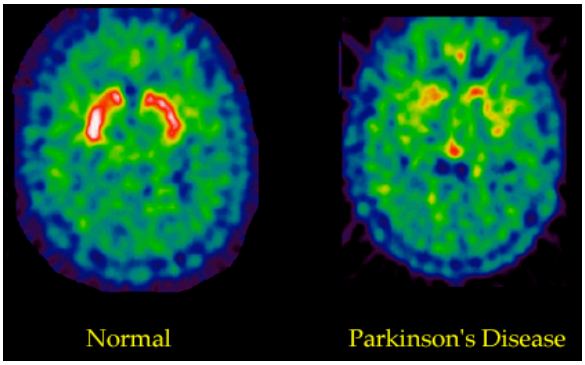
AI Against Aging
Accelerating the Quest for Longevity via Intelligent Software

Dr. Ben Goertzel
CEO, Novamente LLC and Biomind LLC
Director of Research, Singularity Institute for AI
External Research Professor, Xiamen University

Current collaborative work between **Genescient** and **Biomind** involves applying AI tools to understand the genetic differences between **long-lived Methuselah flies** and ordinary flies. This may lead to the discovery of **new pharmaceutical and nutraceutical remedies** for human aging.

AI data analysis done by the University of Virginia and Biomind LLC in 2005 showed that the brain dysfunctions characteristic of **Parkinson's Disease** are rooted in certain patterns of heteroplasmic **mutation in mitochondrial DNA**.



Contents

Aging Is Solvable

There is no longer any reasonable doubt that death due to aging is a solvable problem. The human body is a complex machine, and it is modifiable and reparable like any other machine. A host of researchers, of whom Aubrey de Grey has been the most vocal of late[1], have laid out specific plans for repairing the various interrelated properties of the human body that cause us to age and ultimately die.

But though the roadmap to radical life extension is clearer than ever before, the magnitude of the tasks involved is also apparent. De Grey's plan involves multiple initiatives in seven major areas; each of these initiatives has numerous uncertainties involved with it, and at present there's no way to tell how rapidly any of them will meet with success. All this work, and other related work suggested by other researchers, is of tremendous value and seems very likely to achieve its ends eventually if adequately funded; but one cannot help wondering if there might be some way to accelerate the process. Putting more funding into life extension research as currently practiced and envisioned is critical, but my goal in this article is to suggest a complementary approach based on my own experience as a biomedical researcher and a scientist in other domains.

Radical Life Extension: What Is the Real Bottleneck?

No aspect of the biomedical research pipeline is perfect, so it's a good thing that there are active efforts aimed at improving all of its various aspects, from better experimental machinery to better animal models for testing, etc. However, the weakest link in the pipeline is getting the least attention: this is the effectiveness of the brains of human researchers.

By this I mean no insult to the scientists involved; they are surely some of our species' best and brightest. But the human brain ultimately was not evolved for the integrative analysis of a massive number of complexly-interrelated, high-dimensional biological datasets. We desperately try to cast biological data in a form our human brains can understand effectively: we create data visualizations to ease the application of the 30% of our brain that is customized for visual processing; we develop vocabularies and ontologies to better apply the large portion of our brain customized for linguistics. But there is no portion of the brain customized for generating hypotheses by analyzing biological data. At this stage, the weakest link in the biomedical pipeline is our human brains' lack of ability to holistically understand the mass of data that has been (and is, every day, being) collected, and use this understanding to design new experiments leading to new understanding. This is the primary bottleneck along our path to radical life extension.

Bypassing the Limitations of the Human Brain

There are three evident solutions to this problem: improve the human brain, augment it with external tools, or replace it with something better.

The former is an exciting possibility which will surely be possible at some point, but neuroscience and neuroengineering are currently a long way from enabling robust human cognitive enhancement. Furthermore, advancing neuroengineering is largely a biology problem – which means that a major bottleneck along the path to its achievement is precisely the problem we're talking about, the limitations of the human brain at grappling with masses of biological data.

External tools for biological data analysis are critical and fortunately they are now plentiful, but it is increasingly clear that the sorts of tools we have created are not sufficient to allow us to grapple with the patterns in the data we've collected. Contemporary bioinformatics analysis and visualization software represents a noble, yet ultimately inadequate attempt to work around the shortcomings of the human brain.

To see this, consider the commonplace observation that most geneticists focus their research on a handful of genes, or at very most a handful of biological pathways. This cognitive strategy on the part of researchers makes sense because the human brain can handle only so

much information. There are some genes, p53 for example, about which so much information is known that very few human scientists today have it all in their heads. On the other hand, it's also well known that the human body is a highly complex system whose dynamics are dominated by subtle nonlinear interactions between different genes and pathways. So the correct way to analyze biological data is not to focus on individual genes and pathways, but to take more of a holistic, systems-biology approach.

Can software tools help with this? It turns out the answer is yes – but only to a limited extent. While not as commonly utilized as they should be, there do exist statistical and machine learning approaches to analyzing biological data, which take a holistic approach and extract global patterns from huge datasets. Unfortunately, though, these software programs only go so far; they produce results that still need to be interpreted by human biologists, whose expertise is invariably limited in scope, due to the limitations of human memory.

Visualization tools help a lot here as well, but also have fairly strict limitations: the human eye can only take in so much information at one time. It evolved for scanning the African veldt, not the intricacies of biomolecular systems. Even if you had a holographic simulation of some portion of the human body at some particular scale, this still wouldn't allow the human perceptual system to "see the whole," to grasp all the mathematically evident patterns in the data being visualized.

From a scientific perspective, it would be ideal if we could simply replace human biologists with AI systems customized for biological data analysis – systems with the human capability for insight and interpretation (or even more so), but more memory and more capability for quantitative precision, and pattern-analysis capability tuned for biological data rather than recognizing predators on the veldt. Unfortunately, this kind of "AI scientist" does not exist at present. There are serious research programs underway with the aim of producing this kind of software; and an increasing confidence in the AI field that this is indeed an achievable goal[2]. But life extension research, and biological research in general, cannot afford to wait for computer scientists to produce powerful AI – there's too much urgency about moving ahead with solving medical problems causing human suffering, right now.

Given these realities, my own work in biomedical informatics has focused on a sort of midway point between the approaches of "better tools" and "replace the humans". I am one of those AI scientists who believes that the creation of a powerful AI scientist is a real possibility within the next few decades, maybe even within the next decade. I am involved with two linked enterprises, the commercial AI firm Novamente LLC and the open-source AI initiative OpenCog, oriented specifically toward this goal. But, via my role in the bioinformatics firm Biomind LLC, which is working with the NIH and has worked with the CDC and various academic biomedical labs, I'm also acutely aware that there is important biomedical data being generated right now about important human problems, and we've got to deal with it as best we can. So the approach we're taking is an incremental one: as our ambitious AI scientist is gradually created (and it's a long-term research project, as one might expect), we are utilizing the various modules of the overall AI system to analyze biological datasets. Of course, the AI modules are not as powerful as a full-scale AI scientist would be, but our experience has shown that they can still provide insights beyond what human scientists can achieve unaided, or using conventional tools. In this way AI and biomedical science can progress together: the more progress we make toward the AI scientist, the more powerful the insights generated by the partial versions of the system.

In the rest of this article, I'll describe some work my team at Biomind has done already, applying our AI technology to analyze data regarding aging-related diseases like Parkinson's and Alzheimer's, and potential life extension remedies such as calorie restriction. I will then discuss what I think can be done to push the AI approach to biomedical science further, faster, without waiting for a full-on AI scientist, but using the technologies we have available today. My strong conviction, based on my experience in the AI and bioinformatics fields, is that with a concerted effort to apply current AI tools to biomedical data in a systematic and holistic way, dramatic insights could be achieved, allowing a vast acceleration in the rate of progress of our understanding of aging-related and other diseases, and significantly speeding the advent of radical human life extension.

AI Uncovers the Role of Mitochondrial DNA in Parkinsons and Alzheimers Disease

One of the most exciting chapters so far in my exploration of the application of AI to bioscience, involved work we did in 2005 analyzing data regarding the genetic roots of Parkinson's disease. In this case, the result of the AI analysis was a powerful statistical validation of the hypothesis that Parkinson's is caused by mitochondrial mutations. These results seem reasonably likely to lead to a practical diagnostic test for Parkinson's, and if the work being done at Gencia[3] on protofection works out, they may ultimately form the foundation of a mitochondrial gene therapy based cure.

Over a million Americans have Parkinson's disease. Yet in spite of years of effort by medical researchers, tracking down the genetic roots of the disorder has proved devilishly difficult. The DNA one usually hears about lies in the nucleus of a cell, the cell's center. In many cases the genetic roots of disease can be traced down to mutations in the nuclear DNA, called SNP's or Single-Nucleotide Polymorphisms. Biomind had a significant success with this sort of analysis when analyzing SNP data regarding Chronic Fatigue Syndrome: the AI was able to tease out patterns of mutational combination that provided the first real evidence that CFS is at least partially a genetically-founded disease[4]. While this sort of approach has not proved workable for Parkinson's, a variation proved dramatically successful. Mitochondria, the cell's energy-producing engines, also contain a small amount of DNA. What the AI has told us is that the right place to look for the genetic roots of Parkinson's is in the mutations in the *mitochondrial* DNA. Our software identified a particular region of a particular gene on the mitochondrial genome that appears to be strongly associated with Parkinson's disease[5].

Much smaller, lesser known and lesser studied than its nuclear cousin, the mitochondrial genome is nonetheless vital to cellular function in humans and other animals. The human mitochondrial genome only contains seven genes, whereas the nuclear genome contains around 30,000 at last count. But these seven genes carry out a lot of valuable functions. If they stop working properly, serious problems can arise. In 1999, Dr. Davis Parker, together with Russell H. Swerdlow and scientists from San Diego firm MitoKor's published work suggesting that defects in the mitochondrial genome may be correlated with Parkinson's disease. As a baby's mitochondrial DNA comes entirely from its mother, these results suggest that Parkinson's may be passed maternally -- but that its defects can skip generations, making the emergence of the disease appear random.

The work Parker and Swerdlow's team did involved clever manipulations of embryonic human nerve cells. They removed the mitochondrial DNA from the embryonic nerve cells and replaced it with other DNA: sometimes from healthy people and sometimes from Parkinson's patients. What resulted was the nerve cells receiving the mitochondrial DNA from Parkinson's patients began acting like nerve cells on MPTP. Low complex I activity, meaning insufficient energy obtained from mitochondria – and eventually leading to Parkinson's-like sluggishness.

These results were fascinating and suggestive – but where were the actual mutations? All this showed was that the problem lay somewhere in the mitochondrial genome. The question was where. Which mutations caused the problem?

To answer this question, Parker and colleagues sequenced mitochondrial DNA drawn from the nerve cells of a number of Parkinson's patients, as well as a number of normal individuals, and looked for patterns. But to their surprise, when in 2003 they set about seriously analyzing this data, they found no simple, consistent pattern. There were no specific genetic mutations common to the Parkinson's patients that were not common to samples taken from healthy subjects.

Enter artificial intelligence. Dr. Rafal Smigrodzki, one of Parker's collaborators, was familiar with my AI research work and suggested that perhaps my AI technology might be able to find the patterns in the mitochondrial DNA data.

To make a long story short, it worked. Appropriately enough, the solution turned out to be an AI software technique called "genetic algorithms," which simulates the process of evolution by natural selection – beginning with a population of random solutions to a problem, then

gradually "evolving" better solutions via letting the "fittest" solutions combine with each other to form new ones, and making small "mutations" to the fittest solutions. In this case, what the software was "evolving" was potential patterns distinguishing Parkinson's patients from healthy subjects based on the sequences of amino acids in their mitochondrial DNA. This kind of data analysis is highly exploratory and is never guaranteed to yield a solution – but in this case things worked out happily, and a variety of different data patterns were discovered.

The trick, it turns out, is that while there are no specific mutations corresponding to Parkinson's disease, there are regions – and combinations of regions -- of the mitochondrial genome that tend to be mutated in Parkinson's patients. There are many different rules of the form "If there are mutations in this region of this mitochondrial gene and that region of that mitochondrial gene, then the person probably has Parkinson's disease." While it took some advanced AI technology to find these patterns, once discovered, the patterns are very easy for humans to understand. The patterns were validated by subsequent biological analysis on additional patients[6].

Yet more excitingly, we've done further work (to be published shortly) with Dr. Parker on comparable data regarding Alzheimer's disease, showing patterns that are similar in nature but different in detail. Once again, although the crucial idea to look at the mitochondria in the first place was provided by human biological intuition, the human brain was unable to detect the relevant patterns in the mitochondrial mutation data, even augmented with cutting-edge statistical tools. But AI found the relevant patterns, which are then easily validated via further biological experiments.

AI Helps Unravel the Genetic Mechanisms Underlying the Efficacy of Calorie Restriction for Life Extension

As well as helping to understand and diagnose (and, ultimately, cure) aging-related diseases like Parkinson's and Alzheimer's, AI technology can help us better understand, refine and design methods for extending the maximum lifespan of organisms. One recent example of this is a study my colleagues and I recently published in Rejuvenation Research[7], pertaining to the genetic mechanisms underlying the impact of calorie restriction diets on maximum lifespan. The exact mechanism by which calorie restriction works remains incompletely understood (though there are plenty of theories!), but our AI-based analysis revealed a central role for several genes whose involvement in CR's efficacy was not previously known. These results suggest a number of specific biological experiments, and we are in discussions with biology research labs regarding the best way to carry out these experiments. These experiments of course will produce new data to be analyzed via AI algorithms, and will likely provide information on how various elements of the many existing theories of CR's efficacy combine to provide the true explanation. Through this sort of iterative interaction between AI analysis, human judgment and laboratory experiments, we can progress much faster than would be possible without AI in the picture.

In our application of AI to CR, we initially fed our AI system three datasets that other researchers had posted online, based on their work studying mice on calorie restriction diets. We then merged these three datasets into a single composite dataset for the purpose of conducting a broader-based analysis, using AI technology rather than the standard statistical methods that the researchers had originally used on their datasets.

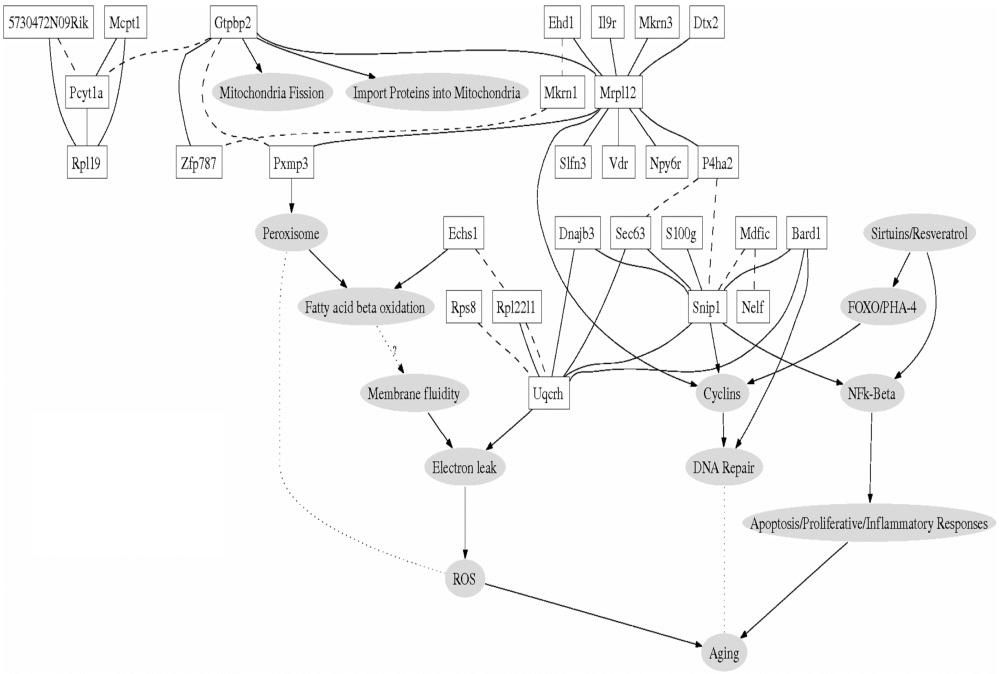
Along with providing a large amount of other information, this analysis resulted in a list of genes that the AI found to be important for CR's impact on lifespan. An essential point here is that the AI was capable of teasing out nonlinear interactions between different genes and gene products. The genes that the AI points out as important for CR and its impact on aging are important, not necessarily in terms of their individual actions, but most often largely in terms of their interactions with other genes.

The AI also provided a map of gene interrelationships (shown in Figure 1), suggesting which *inter-gene interactions* are most important for the effect of CR on life extension. In particular, our graphical analysis revealed that the genes Mrpl12, Uqcrh and Snip1 play central

roles in the effects of CR on life extension, interacting with many other genes (which the analysis enumerates) in carrying out their roles. This is the first time that the genes Snip1 and Mrpl12 have been identified as important in the aging context.

To double-check the validity of these results we obtained from analyzing three datasets at once, we then ran the same AI processes all over again, but throwing a fourth dataset into the mix. Much to our relief the results were largely the same – suggesting that the AI is producing real biological insights, not just some kind of data processing artifacts.

Broadly, the biological interpretation of these analytical results suggests that the effects of CR on life extension are due to multiple factors, including factors identified in prior theories of aging, such as the hormesis[8], development[9], cellular[10] and free radical[11] theories. None of these individual theories stands out as obviously correct, based on the patterns of gene-combination effects identified by the AI system. But genes with predicted involvement according to many of these theories play a role, along with other genes not highlighted by any prior theories or experiments.



Unraveling the Mystery of the Methuselah Flies

One of our current research projects involving AI and biology data has to do with the Methuselah flies: fruit flies that have been bred by directed evolution, over the last 30 years, to live 5x longer than ordinary fruit flies. Simply by setting up a situation where longer-lived flies are more likely to breed with each other, and letting it operate for many many generations, a new strain of flies was created. This is a miracle, and a puzzle: becaues these Methuselah flies were created via directed evolution rather than genetic engineering or other “direct” methods, we don’t know how what it is that makes them live so long. Now we need to “reverse engineer” what directed evolution did, and understand what combination of genetic mutations occurred to create the long-lived flies, and why these mutations had the impact they did. This is not a completely simple matter, because evolution is messy: the Methuselah flies are bound to have a lot of inessential differences from regular flies along with the functionally critical ones, and the

inessential and critical ones are going to be complexly bound up with each other. Traditional statistical analysis methods can identify some genes that are important to understanding the difference between the Methuselah flies and ordinary flies, but, they can't unravel the genomic, proteomic and metabolomic interrelationships.

Even without a full understanding of what keeps them ticking, analysis of the Methuselah flies has borne some fruit (sorry!). Genescient, the company that now has rights to the IP of the Methuselah flies, has used the Methuselah flies to find some substances that can be fed to normal flies to make them live much longer than usual. Furthermore, this research has led to insights regarding nutraceuticals for promoting longevity in humans. But these results are minor compared to what could be achieved if the essential cause of the Methuselah flies' longevity were understood. Not all biologists agree that understanding aging in fruit flies will help us understand human aging, but there's a strong argument to be made. To the extent that aging is a basic property of cellular function, it is likely to be the same process across many organisms – and indeed, Genescient has done studies showing that the genes most significant in characterizing the Methuselah flies tend to be ones that also relate to human diseases.

Aubrey de Grey's "engineering approach" to combating aging focuses on the symptoms of aging, which occur at various levels throughout the body (as a single example, he proposes to use certain bacteria to clean up "gunk" that appears between the body's cells increasingly with age). While this approach may have great value, there's also something to be said for trying to fix the basic cellular processes underlying aging. Perhaps if these processes are fixed then many of the symptoms will disappear on their own. Ultimately, de Grey's approach and Genescient's approach may lead to complementary therapies.

So far we have only applied AI to a small set of fly data, but we have already found some interesting conclusions. The general role of AI here is to identify which genes are important for the Methuselah flies' longevity, and how these genes combine with each other -- and based on this understanding, to figure out which pathways can be impacted with pharmaceuticals or nutraceuticals to cause ordinary flies to live longer. AI can also select among the various relevant genes and pathways to estimate which ones are most likely to lead to human aging therapies. As in our previous examples, the AI is far from autonomous here; it is serving as a helper to human biologists and data analysis. But there is a lot of data and the biology is complex, so the latter can use all the help they can get!

I can't recount the details results of our work with Genesicent here due to intellectual property concerns, but I can review the basic sorts of things we're finding. For instance, we have found one gene that seems to be very important to fly longevity, and that produces a certain enzyme known in humans due to its deficiency in people with a certain monogenetic disease involving central nervous system malfunction. Another gene emerging as important is a tumor suppressor gene (and the relation between cancer tumor suppression and aging is very well known), which plays a role in the Methuselah flies in combination with several particular genes related to metabolism. None of these findings, in itself, tells you why the Methuselah flies live so long – but they point research in specific directions, some of which would not have been conceived based on this data without the AI-based analysis results.

AI That Reads Biological Texts

So far we've been discussing the use of AI to analyze quantitative biological datasets. But there's another fact that must also be considered, which is that the vast majority of biomedical knowledge online right now exists only in textual format. Most datasets aren't placed online, and as big as the biological databases are, most knowledge that could be placed in there, actually hasn't been, either because no one has gotten around to it, or because researchers prefer to keep their data proprietary.

For example, at Biomind we've done a lot of work with the Gene Ontology, which is an outstanding database that categorizes genes by function. If you look up "apoptosis" in the Gene Ontology, you'll find a few dozen genes that have been categorized as being associated with apoptosis--preprogrammed cell death. But the catch is, if you browse through the journal

literature online, you'll find even more. The Gene Ontology can't keep up. This is a tribute to the rapid pace of biomedical research these days, but it's also an indication of one direction biomedical software has got to go in: We've got to write computer programs that can grab the information directly from the texts where it's been published! This is a domain of research called Bio-NLP – bio natural language processing.

Once a sufficiently powerful AI scientist is created, Bio-NLP won't be necessary, as the AI will simply recognize all the relevant patterns in the data directly, without need for human insight. But we're not there yet. So at the present time, the best strategy for AI data analysis is to incorporate all available sources of information, including direct experimental data and text humans have produced based on interpreting that data.

In 2006, I co-organized the sixth annual Bio-NLP workshop, as part of the annual HTL-NAACL Computational Linguistics conference. At previous Bio-NLP workshops, nearly all the work presented had pertained to fairly simple problems, such as recognizing gene and protein names in research papers (a task made more difficult than it should be by the presence of multiple naming conventions among biologists). But starting in 2006 we saw more and more researchers creating software with the capability to recognize *relationships* between biological entities, as expressed in natural language text; and this trend has intensified subsequently. The latest Bio-NLP software (see Rzhetsky's work[12] for an impressive example) takes in a research paper and tells you which genes, proteins, chemical and pathways are mentioned, and how they are proposed by the authors to relate to each other (which genes are in which pathways, which enzymes catalyze which reactions, which genes upregulate which others, etc.). This is a far cry from full understanding of the contents of research papers, but it's definitely a start.

Premise 1	Importantly, bone loss was almost completely prevented by p38 MAPK inhibition. (PID 16447221)
Premise 2	Thus, our results identify DLC as a novel inhibitor of the p38 pathway and provide a molecular mechanism by which cAMP suppresses p38 activation and promotes apoptosis. (PID 16449637)
(Uncertain) Conclusions	DLC prevents bone loss. cAMP prevents bone loss.

AI Based Logical Inference Based on Information Automatically Extracted from PubMed Abstracts

The paper I presented at Bio-NLP 2006 regarded a research prototype called BioLiterate, which we built for the NIH Clinical Center in 2005. What the BioLiterate prototype did was extract relationships from various biomedical research abstracts, and try to glue them together using logical reasoning. So, for example, if one paper said that p38 map kinase inhibition prevents bone loss, and another paper said the DLC inhibits p38, then the software would put A and B together, deciding (using logical reasoning) that maybe DLC prevents bone loss (the actual sentences the AI used in these inferences, found in PubMed abstracts, are shown in the figure above). The logical inference was provided by the Probabilistic Logic Networks module of the Novamente Cognition Engine[13]. BioLiterate was a prototype, rather than a robust and deployable software solution, but it made its point: If you build a Bio-NLP system and then use the right sort

of rules to pipe its output into a computational reasoning system, you get an automated biological hypothesis making system.

A Bolder Approach: The Holistic Biobase

The work we've done so far, applying AI to bioinformatics, has already led to exciting results. Continuing the approach, applying AI technology to various datasets in isolation, there is little doubt that an ongoing stream of comparable results can be obtained, providing a significant and worthwhile acceleration to the advancement of bioscience.

But, we could do a lot better. The real future of bioscience, I am convinced, lies in the simultaneous analysis of a lot more than the four datasets we considered in our calorie restriction study. We need to feed dozens, hundreds, thousands, tens and hundreds of thousands of datasets simultaneously into the same AI system – along with all the biological texts online -- and let the AI go to town hunting down the patterns that are concealed therein. AI can detect far more patterns in such a data-store than the human mind.

Right now, the mass of available data is terrifyingly underutilized, due to the limitations of the human brain and the corresponding processes of the scientific community (which are adapted to the limitations of the human brain). Human scientists analyze individual datasets, or small collections of datasets, using brains that evolved for solving other sorts of problems (aided by statistical, visualization and in rare cases AI tools); and then these humans write papers summarizing their results. Of course, the papers written about a certain dataset ignore nearly all the information in that dataset, focusing on the particular patterns that the researchers noticed (which are often the ones they were looking for in the first place, based on their prior knowledge and biases). Then, researchers read the papers other researchers have written, and use the conclusions in these papers to guide the analysis of new datasets. The multiple datasets that have been collected are brought together indirectly only via human beings reading and writing papers, each of which contains an extremely partial view into the data on which it's based. This is a dramatic, tragic loss of information compared to what would happen if the datasets were actually analyzed collectively in a serious way.

What I am suggesting is that we create a ***Holistic Biobase*** – a massive data repository containing all the biomedical information on the Web today – including quantitative data, relational data, textual information in articles and abstracts ... everything. The data in this repository should then be analyzed using powerful AI systems that are able to study the data as a whole, identifying complex patterns not amenable to direct human analysis nor conventional statistics. These software systems will help humans make better discoveries, and in some cases they will surely make new discoveries on their own – suggest new experiments, propose new hypotheses, make connections that no human could make due to our limited ability to store and analyze information in our brains.

The Holistic Biobase should ideally be an open information resource, so that any scientist with statistical or AI tools and a bit of savvy can crunch the data in their own way. A decent, if partial, model for the Holistic Biobase is Freebase[14], which is an open online database containing various sorts of information of general interest. In principle, one could just load biological datasets into Freebase, but in practice this isn't likely to be the best approach, for several reasons. Freebase is a traditional relational database, which is not the most natural data structure for AI purposes (a graph database would be preferable). And more critically, it doesn't solve the problems of metadata standardization and data normalization, which are perhaps the main obstacles standing in the way of constructing the variety of mega-bio-database I'm envisioning.

If the Holistic Biobase concept sounds overambitious and fanciful, please remember that the Human Genome Project once sounded very much the same. A few decades ago the "synthetic organism" project of Venter's lab at the J. Craig Venter Institute in Rockville, Maryland would also have sounded science-fictionally speculative. And how many people would have labeled the notion of a Google-scale database of online documents implausible or insane, just

one or two decades ago? Biology and computer science both are in the midst of phases of rapid advance, which opens up possibilities that could barely have been conceived of before.

As a very simple example of the value the Holistic Biobase would have, let's turn back to the calorie restriction data analysis project mentioned above. We're excited with the results we achieved based on our four-dataset analysis – but it's easy to see how much more powerful the results could be if we had a massive integrative data repository at our disposal. For example, calorie restriction is connected with energy metabolism, a connection we as humans can exploit by interpreting the results of calorie restriction data in the context of our own knowledge about energy metabolism pathways. But what if we integrated masses of raw data regarding energy metabolism in various aging-related contexts into the analysis – and looked at this data together with the calorie restriction data? Who knows what might turn up? Bodies are complex systems, and the effect of calorie restriction on life extension is surely not a phenomenon best understood in isolation. And of course there are a dozen other pathways that should be considered along with energy metabolism.

What kinds of AI algorithms will be able to grapple with the Holistic Biobase in a really effective way? We don't have much experience doing this kind of massive-scale biological data analysis, but the experience we do have gives us significant guidance. There have already been some commercial products pushing in this direction – for instance Silicon Genetics' GeNet database (for microarray data) and associated MetaMine statistical datamining package. But GeNet/Metamine handles only standard statistical methods, and applies only to microarray data. On the other hand, the methods we've been using in Biomind to date are more advanced analytically and are oriented toward combined analysis of multiple types of data. However, they have not yet been tailored for massive-scale data analysis.

My strong suspicion is that to handle the Holistic Biobase, new methods will be needed. Current applications of AI to bioinformatics have focused on the application of machine learning algorithms for pattern recognition – essentially, algorithms that look at one or more datasets and explicitly scan them for patterns using complex algorithms. To handle larger numbers of data and yet preserve the capability for analytical sophistication, a paradigm shift will be required – and this paradigm shift ties in naturally with the trends of development in the AI field itself. What is needed is the fusion of bioinformatics data analysis with *automated reasoning*. More specifically: automated *probabilistic* reasoning, since biological data is riddled with uncertainties. Automated reasoning allows an AI system to study a handful of datasets, derive results regarding the patterns in these datasets, and then extrapolate these patterns to see what they imply about other datasets. This step of inferential extrapolation allows far more scalable analysis than machine-learning pattern-recognition methods alone. My own team is currently pursuing this vision via integrating our OpenBiomind bioinformatic AI software with our OpenCog general-AI platform, which includes a powerful probabilistic inference framework called Probabilistic Logic Networks[15]. This will allow us, for example, to massively extend our calorie restriction data analysis project, to include numerous datasets drawn from studies of different but allied biological phenomena.

While this vision goes a fair bit beyond current practice, there are some contemporary projects with smaller but vaguely similar ambitions. One example is a project called ImmPort – this is a program funded by the National Institute of Health, specifically the National Institute for Allergies and Infectious Diseases, which Biomind is involved with via a subcontract to Northrop-Grumman IT. ImmPort is still in the making, but what it's going to be, when it's finished, is a Web portal site for NIH-funded immunologists. Biomind's role has been to integrate bioinformatics analysis technologies into the portal, both our own innovative machine learning techniques and more standard methods. The most exciting part of ImmPort is probably its potential to enable massive data integration. When an immunologist uploads data into ImmPort, it will automatically be put in a standard format, so it can be automatically analyzed in the same way as all the other datasets that were uploaded – and, most excitingly, so it can be analyzed in terms of the patterns that emerge when you put it together with all the other datasets. This is something that's hardly ever being done right now – the application of bioinformatic technology to look for patterns spanning dozens or hundreds or thousands of datasets. However, the scope of AI currently envisioned within ImmPort is restricted to machine-learning algorithms; extension to more powerful automated inference methods is beyond the scope of the project.

Projects like ImmPort are definitely a step in the right direction – but only a step. Even if every immunologist on the planet were to upload their data into ImmPort, and even if ImmPort were to incorporate inference-based data analysis, the restriction to immunological data alone would still constitute a huge limitation. The immune system is not an island, it is intricately connected with nearly all other body systems. As an example, in our work with the CDC we found that CFS is most likely a complex interaction between immune, endocrine, autonomic nervous and other functions. What we need is not just a holistic immunology database but a holistic biology database, and with a focus on powerful cross-dataset AI analysis as well as statistical and machine learning methods. Furthermore, there is as yet no ImmPort analogue for data directly related to life extension.

My hope is that over the next few years the ideas in this article will become boring and mainstream, and the value of massive, sophisticated, AI-based cross-dataset analysis will move from outrageous to obvious in the consensus view. Until that time, we will continue to be slowed down in our quest to extend human life and cure human disease by the limitations of our human brains at analyzing the relevant biological data.

References

[1] De Grey A, Rae M. Ending Aging: The Rejuvenation Breakthroughs That Could Reverse Human Aging in Our Lifetime. St. Martin's Press 2007.

[2] Goertzel BN, Coelho LS, Pennachin C, Goertzel I, Queiroz M, Prosdocimi F, Lobo F. Learning Comprehensible Classification Rules from Gene Expression Data Using Genetic Programming and Biological Ontologies. In: 7th International FLINS Conference on Applied Artificial Intelligence: 2006; Genova, Italy; 2006.

[3] Smigrodzki RM, Khan SM. Mitochondrial microheteroplasmy and a theory of aging and age-related disease. Rejuvenation Res. 2005 Fall;8(3):172-98.

[4] Goertzel BN, Pennachin C, de Souza Coelho L, Maloney EM, Jones JF, Gurbaxani B. Allostatic load is associated with symptoms in chronic fatigue syndrome patients. Pharmacogenomics 2006; 7:485-494.

[5] Smigrodzki R, Goertzel B, Pennachin C, Coelho L, Prosdocimi F, Parker WD, Jr. Genetic algorithm for analysis of mutations in Parkinson's disease. Artificial Intelligence in Medicine 2005; 35:227-241.

[6] Smigrodzki R, Goertzel B, Pennachin C, Coelho L, Prosdocimi F, Parker WD, Jr. Genetic algorithm for analysis of mutations in Parkinson's disease. Artificial Intelligence in Medicine 2005; 35:227-241.

[7] Goertzel B, Coelho L, Mudado M. "Identifying the Genes and Genetic Interrelationships Underlying the Impact of Calorie Restriction on Maximum Lifespan: An Artificial Intelligence Based Approach", to appear in Rejuvenation Research.

[8] Sinclair DA. Toward a unified theory of caloric restriction and longevity regulation. Mech Ageing Dev 2005; 126:987-1002.

[9] de Magalhaes JP, Church GM. Genomes optimize reproduction: aging as a consequence of the developmental program. Physiology (Bethesda) 2005; 20:252-259.

[10] Shay JW, Wright WE. Hallmarks of telomeres in ageing research. J Pathol 2007; 211:114-123.

[11] Harman D. Aging: a theory based on free radical and radiation chemistry. J Gerontol 1956; 11:298-300.

[12] Krauthammer M, Rzhetsky A. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. Bioinformatics 2001;17(Suppl 1):S74–82 (0)

[13] Ikle M, Goertzel B, Goertzel I, Heljakka A. "Probabilistic Logic Networks: A Comprehensive Framework for Uncertain Inference", Springer, May 2008.

[14] Freebase: an open, shared database of the world's knowledge. http://www.freebase.com

[15] Ikle M, Goertzel B, Goertzel I, Heljakka A. "Probabilistic Logic Networks: A Comprehensive Framework for Uncertain Inference", Springer, May 2008.