Machine Learning and Bioinformatics

Slide 1

Bioinformatics is the field where biology meets computer science and statistics. It's about making sense of massive biological data — DNA sequences, proteins, medical records — data that is far too big and complex for humans to process alone.

BLAST (Basic Local Alignment Search Tool)

The story really began with early computational tools. Before machine learning, one of the foundational tools in bioinformatics was **BLAST**, which stands for *Basic Local Alignment Search Tool*. It allowed researchers to compare a new DNA or protein sequence against huge databases and find regions of similarity. It worked using clever scoring matrices and heuristics to detect matches quickly.
This was revolutionary in the 1990s and early 2000s — suddenly scientists could identify genes, study evolution, or find potential drug targets in silico. But the challenge was that BLAST required **trained scientists** to interpret the results carefully. It wasn't scalable for the kind of massive, noisy data we generate today, like whole genomes or medical imaging. That's exactly where machine learning has stepped in: to handle scale, pattern complexity, and even predictions that BLAST could never make.

Human Genome Project (HGP)

Another major milestone in bioinformatics was the Human Genome Project, completed in 2003. It was the first time scientists mapped the entire human DNA sequence — over 3 billion base pairs. But it came at an enormous cost: billions of dollars and more than a decade of work.

The HGP gave us a reference genome, but analyzing and comparing it was still painfully slow. Traditional tools like BLAST could help, but they weren't scalable enough for population-level studies. The result was that while we had the data, extracting actionable insights — like disease risks or drug targets — was limited by our algorithms and computing power.

Today, thanks to the advances in sequencing and ML, what once took years and global collaborations can now be done in weeks or even days, allowing us to move from simply sequencing genomes to actually interpreting them for health and medicine.

But the real challenge isn't sequencing anymore — it's analyzing this enormous flood of data. Traditional methods hit their limits. They could store data and align sequences, but they struggled to uncover hidden patterns — like how a specific mutation might cause cancer, or how proteins fold into complex structures.

This is where machine learning changes the game. Instead of writing explicit rules, ML models learn patterns directly from data. With enough data, they can classify genes, predict diseases, and even discover new drugs. From BLAST to AlphaFold, the field has moved from static databases to dynamic learning systems, powered by machine learning.

(Diagram Difference between HGP and ML)

Slide 2

One of the biggest applications of machine learning in bioinformatics is disease diagnosis.

One such example is Google Health's diabetic retinopathy project. Retinopathy is a diabetes-related eye disease that can cause blindness if not caught early, it is a leading cause of preventable blindness. Screening traditionally requires trained ophthalmologists, which limits access in many parts of the world. Google developed deep learning models that analyze retinal images and match or even exceed human doctors in detecting early signs. This is a real-world case where ML has directly improved healthcare accessibility.

Another example comes from Stanford University, where researchers trained convolutional neural networks on over 120,000 images of skin lesions. Their AI system performed on par with dermatologists in identifying skin cancers. This is a big deal because skin cancer is one of the most common cancers, and accurate early detection saves lives.

These are not just research prototypes — they are being tested in real-world clinical settings.

Genomics is another powerful area. Instead of looking at medical images, ML models can analyze DNA sequences to predict disease risk. One example is DNABERT — a model that applies transformer architecture, originally used in language models, to DNA sequences. It learns the "language of the genome" and can predict mutations linked to diseases. Similarly, BioBERT was trained on biomedical literature, helping researchers mine vast databases of medical papers to find disease-gene associations quickly. The benefit here is speed, accuracy, and scalability. ML can look at millions of sequences or thousands of patient records in seconds, uncovering connections that would take humans years. Of course, there are challenges — bias in datasets, overfitting, and interpretability. A model might be accurate but give no explanation for its decision, which makes doctors hesitant to trust it. Still, progress in explainable AI is helping bridge this gap. Overall, ML is moving from the lab into the clinic, and it's already saving lives by making diagnoses faster and more reliable.

Slide 3

Beyond diagnosis, machine learning is transforming drug discovery — a process that has historically been slow and extremely expensive. On average, developing a new drug can take more than 10 years and billions of dollars. One of the biggest breakthroughs came from DeepMind's AlphaFold, which I'd argue is one of the biggest breakthroughs in biology in decades, published in 2021. AlphaFold uses deep learning to predict protein structures with astonishing accuracy. Why does this matter? Proteins are the building blocks of life, and their shape determines their function. If we can predict a protein's 3D structure from its DNA sequence, we can design drugs that interact with it precisely. This was once considered one of the grand challenges of science. Experimental methods like X-ray crystallography took months or years per protein. AlphaFold's deep learning approach solved this in hours, producing results nearly as accurate as lab experiments. Suddenly, we had structural data on hundreds of thousands of proteins, massively accelerating our ability to understand diseases and discover new drugs.

Models like ProteinBERT extend this approach, treating proteins like sentences and learning their function from sequence data. These models open up possibilities for designing new molecules that could become drugs. In drug discovery pipelines, reinforcement learning and generative models are now used to design novel drug candidates.

Companies like Insilico Medicine and BenevolentAI are already applying ML to accelerate R&D. For example, Insilico Medicine used AI to identify a new drug candidate for idiopathic pulmonary fibrosis in less than 18 months — a process that would normally take years.

Looking ahead, machine learning is driving us toward personalized medicine — tailoring treatments based on an individual's genetic profile. Imagine starting with a patient's genome and predicting which drugs will work best, or even designing new drugs specifically for that person. This isn't science fiction.

One such example is in CRISPR gene editing. Microsoft researchers developed models called Azimuth and Elevation. Azimuth predicts the on-target efficiency of CRISPR edits — in other words, how well the edit will work. Elevation predicts the off-target risks — potential unintended edits elsewhere in the genome. By combining these, scientists can design safer and more effective gene-editing experiments. This kind of predictive power was impossible before ML.

Conclusion

To conclude, bioinformatics started with simple rule-based tools and huge data projects like the Human Genome Project. But the real transformation came with machine learning — from diagnosing diseases using DNA and images, to predicting protein structures, to designing new drugs. We've moved from storing biological data to understanding it — and machine learning is the key that makes this possible.