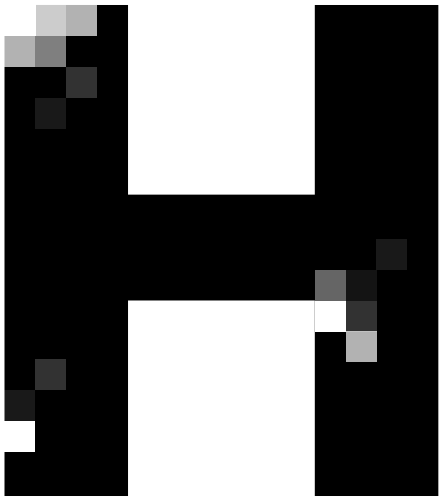
Superdatorer som denna i Linköping, öppna databaser och allt bättre algoritmer har banat vägen för AI, som nu kan lista ut hur proteiner veckar sig i avancerade 3D-strukturer.

AI avslöjar protein-strukturer på nolltid

Text Maria Selmer Foto Thor Balkhed

På några få timmar kan AI ta fram proteinstrukturer som annars skulle ta flera månader, eller till och med år, att få fram. Det innebär ett paradigmskifte för proteinforskningen. Men den nya tekniken kan inte ersätta alla experiment.



ela forskarvärlden häpnade i november 2020. Med hjälp av artificiell intelligens, AI, hade Deepmind med sitt program Alphafold 2 lyckats förutsäga proteinstrukturer från sekvenser med så hög träffsäkerhet att ledningen för den återkommande tävlingen CASP, Critical assessment of protein structure prediction, misstänkte att de hemlighållna experimentella strukturerna hade läckts.

Så var inte fallet. Programmet var bara ett stort steg före alla andra – och hade lyckats förutsäga strukturmodeller vars likhet med experimentella strukturer var på samma nivå som mellan olika strukturer av samma protein bestämda med etablerade metoder. Sommaren 2021 släpptes programvaran fri och genom ett samarbete med institutet European bioinformatics institute, EBI, blev en databas med strukturmodeller av alla proteiner från människa och 20 andra modellorganismer tillgängliga. Världens forskare kan nu snabbt och med högre träffsäkerhet än någonsin tidigare få en bild av hur varje enskilt protein ser ut och därmed få nya ledtrådar till molekylära mekanismer.

DEN GENETISKA KODENS översättning till proteinets aminosyrasekvens har varit känd sedan 1960-talet. Men det har varit en gåta för forskarna hur aminosyrasekvensen sedan bestämmer hur den tredimensionella strukturen hos det funktionella proteinet blir. Det första proteinet – myoglobin från kaskelottval – strukturbestämdes redan 1959. Sedan dess har forskare med hjälp av experimentella metoder, såsom kristallografi, kärnmagnetisk resonans, NMR och elektronmikroskopi, strukturbestämt mer än 70 000 unika proteiner och de komplex de bildar med småmolekyler, nukleinsyror och andra proteiner.

Proteinstrukturerna som tas fram med AI är mycket bra, men inte perfekta. Programmen ger en uppskattning av noggrannheten och kvaliteten hos förutsägelsen för olika delar av modellen. Den är ofta högre för enskilda välstrukturerade delar, domäner, av ett protein och lägre för mer flexibla delar samt för hur de olika domänerna är positionerade. Noggrannheten är också något lägre vid förutsägelse av strukturer med väldigt unik sekvens, som därför sämre matchar det som redan finns i databaserna.

Genombrotten i strukturförutsägelse får omedelbart stora konsekvenser i många grundvetenskapliga och tillämpade forskningsområden. Strukturbaserad läkemedelsutveckling slog igenom i början av 1990-talet, då bland annat hiv-proteasinhibitorer utvecklades med hjälp av proteinets struktur. Området har sedan dess utvecklats till ett stort forskningsfält med ett fruktbart samspel mellan beräkningsbaserade och experimentella metoder. Kunskap om strukturen hos coronavirusets spikprotein har varit viktig för att kunna ta fram vacciner mot viruset, och strukturbiologiska studier av livets grundläggande processer har legat till grund för många Nobelpris.

DE PROTEINSTRUKTURER SOM kan förutsägas med AI-algoritmer ger en ny startpunkt för alla dessa forskningsprojekt. För proteinfamiljer där inga experimentella strukturer finns, lägger de AI-beräknade strukturerna grund för hypoteser som går att testa experimentellt. Strukturmodellerna kan användas för simulering av hur småmolekyler, exempelvis läkemedelskandidater, kan binda till ett enzym eller en receptor. Modellerna kan också användas för att lösa kristallstrukturer eller tolka lågupplösta elektronmikroskopibilder av proteinkomplex. Storskalig strukturförutsägelse applicerat på mängder av proteiner från ett stort antal organismer kommer att vara mycket värdefullt inom bioteknik, för att ge en bättre grund till att välja kandidatproteiner till olika tillämpningar. Inom läkemedelsforskningen kan dessa strukturmodeller ge viktig information, både för läkemedelsmål och möjliga sidoeffekter.

AI gör det möjligt att använda all tillgänglig information i form av experimentellt bestämda strukturer och kända proteinsekvenser. Genombrottet skulle alltså över huvud taget inte varit möjligt utan både utveckling av AI-algoritmer och tillgänglig information från vilken algoritmerna kan lära sig samband mellan proteinsekvens och proteinstruktur.

I nuläget är AI-metoder därför bäst på att förutsäga välstrukturerade delar av proteiner, och proteiner som har någon form av likhet med andra, vars struktur är känd.

Många av proteinernas livsuppehållande funktioner i cellen bygger på interaktioner med andra proteiner, men även inom detta område har AI-metoder gjort framsteg. Andra funktioner är intimt kopplade till interaktioner med nukleinsyror eller småmolekyler. Detta är i nuläget utmaningar för AI-algoritmerna i standardkonfiguration och kräver vidareutveckling. För vissa av dessa tillämpningar är dessutom mängden kända strukturer sannolikt begränsande för vad som går att åstadkomma med maskininlärning.

AI-BASERADE strukturförutsägelser har flyttat fram startlinjen, men de mest intressanta aspekterna kommer fortfarande under en överskådlig framtid att kräva experimentella studier. Vi behöver utforska på vilka sätt många proteiner ändrar struktur när de binder till ligander eller andra makromolekyler, och hur proteiner bildar stora funktionella komplex. Vi kan inte heller i detalj förutsäga de interaktioner som förklarar varför exempelvis en inhibitor binder starkare till ett protein än en annan. Vi har precis börjat processen att till fullo utnyttja nuvarande AI-baserade verktyg, och sannolikt kommer nya AI-metoder i framtiden att flytta fram gränserna för vad som kan förutsägas.

En annan frågeställning som kvarstår är den biofysikaliska processen – hur proteiner i en levande cell bildar sina definie-

Programmet använder information i öppna databaser

Steg 1

```
MKAKSNNYRGKVDISVSQNFITSKNT
IYKLIKKTNISKNDFVIEIGPGKGHI
TEALCEKSYWVTAIELDRSLYGNLIN
KFKSKNNVTLINKDFLNWKLPKKREY
KVFSNIPFYITTKIIKKLLLEELNSP
TDMWLVMEKGSAKRFMGIPRESKLSL
LLKTKFDIKIVHYFNREDFHPMPSVD
CVLVYFKRKYKYDISKDEWNEYTSFI
SKSINNLRDVFTKNQIHAVIKYLGIN
LNNISEVSYNDWIQLFRYKQKID
```

AMINOSYRASEKVENS

Aminosyrasekvens för ErmQ, ett enzym med ännu så länge okänd struktur. Varje bokstav motsvarar en aminosyra. ErmQ metylerar bakteriers ribosomer, vilket gör att antibiotikans bindningsställe blockeras och bakterierna blir resistenta mot antibiotikan erytromycin.

rade tredimensionella strukturer. Trots att vi med hög träffsäkerhet kan förutsäga slutresultatet kan vi i nuläget bara simulera vägen dit för ganska små proteiner. Proteiners veckningsprocess, dynamik och aggregering är nu viktiga utmaningar för beräkningsbaserade metoder.

”De mest intressanta aspekterna kommer fortfarande att kräva experimentella studier.”

Alphafold 2 är ett utmärkt exempel på hur AI kan flytta fram gränserna för hur vi utnyttjar all tillgänglig information för att fatta beslut. Modellerna hjälper forskare till mer välgrundade hypoteser som kan fokusera forskningen där den gör störst nytta. Inom strukturbiologin kan vi med hjälp av experiment och beräkningar ta oss an de frågeställningar där våra insatser verkligen behövs, för att förstå biologiska samband mellan sekvens, struktur och funktion, och för att utveckla tillämpningar inom bioteknik och medicin.

Maria Selmer är professor i strukturbiologi vid institutionen för cell- och molekylärbiologi, Uppsala universitet. Hennes forskning handlar om proteinsyntes, antibiotikaresistens och evolution.

Så fungerar Alphafold 2

I princip alla proteinsekvenser och proteinstrukturer som publiceras finns tillgängliga i öppna databaser, som Universal protein resource, Uniprot, och Worldwide protein data bank (PDB). Med hjälp av PDB, som i år firar 50-årsjubileum, kan alla som vill söka upp, analysera och ladda ned atomkoordinater för makromolekyler med känd struktur.

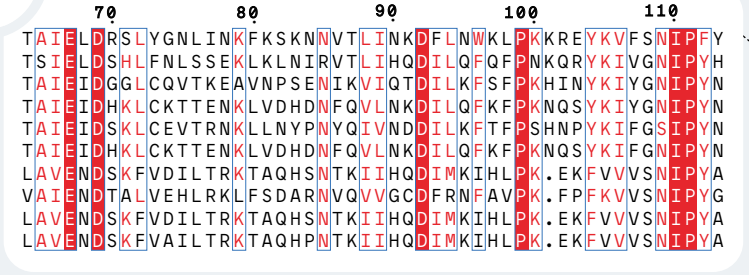
Alphafold 2 utnyttjar alla dessa öppet tillgängliga proteinstrukturer i en maskininlärningsalgoritm. Dessutom utnyttjar programmet information om i vilken utsträckning aminosyror i olika positioner i sekvensen är evolutionärt konserverade, det vill säga om samma eller en liknande aminosyra förekommer i motsvarande position i besläktade proteiner.

Från databaserna extraheras samband mellan aminosyrasekvens och övergripande struktur, men också information om hur aminosyrornas sidokedjor packas optimalt i proteiner. De utgångspunkterna har använts även i tidigare algoritmer för att förutsäga proteinstrukturer.

En anledning till att genombrottet kommer nu är att vi nu har tillgång till strukturer av de flesta typer av domäner, de enskilda veckade delarna av proteinerna. Den andra anledningen är att AI-algoritmer på ett bättre sätt kan utnyttja informationen i stora datamängder, och därför i stort sett hoppa över biofysikaliska antaganden.

Med utgångspunkt från samma data har en alternativ maskininlärningsalgoritm, Rose TTA Fold, utvecklats av forskare vid University of Washington. Denna presterar nästan lika bra och blev öppet tillgänglig i somras.

Steg 2A



SEKVENSDATABASER
Aminosyrasekvensen jämförs med kända sekvenser i databaser. Sekvenser som liknar ErmQ:s plockas ut och sammanställs till en så kallad sekvensupplinjering, vilken visar i vilka positioner som det alltid är samma eller liknande aminosyror (markerat i rött).

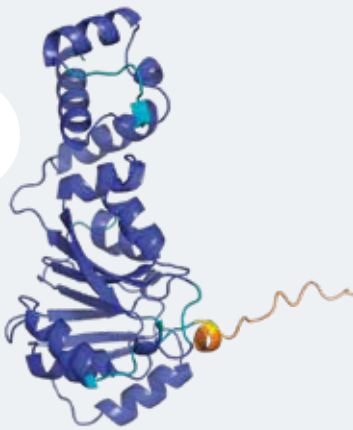
Steg 2B

STRUKTURDATABASER
Aminosyrasekvensen jämförs även med sekvenser för vilka proteinets struktur är känd. Det bidrar med information om samband mellan aminosyrasekvens och tredimensionell struktur hos proteiner, och om hur olika kemiska grupper brukar packas inuti proteiner.

Steg 3

FÖRFINING
Förutsägelsen förfinas ytterligare med hjälp av information från sekvens- och strukturdatabaserna.

Steg 4



STRUKTURMODELL
Strukturmodell av ErmQ framtagen med Alphafold 2. Strukturen har en färgskala där blått visar delar som bestämts med mycket hög säkerhet. Orange färg visar delar med stor osäkerhet.