Polyproliini II -sekundaarirakenteen ennustaminen neuroverkoilla

Markku Siermala

Tampereen yliopisto
Tietojenkäsittelyopin laitos
Pro gradu -tutkielma
Syyskuu 1999

Tampereen yliopisto
Tietojenkäsittelyopin laitos
SIERMALA MARKKU: Polyproliini II-sekundaarirakenteen ennustaminen neuroverkoilla
Pro gradu -tutkielma 78 sivua, 14 liitesivua
Syyskuu 1999

Tiivistelmä

Tutkimuksen kohteena on proteiineissa esiintyvän harvinaisen sekundaarirakenteen polyproliini II (PPII) ennustaminen neuroverkolla. Neuroverkon opetusaineisto haettiin Protein Data Bank -tietokannasta. Tietokannasta poistettiin kaikkiaan 6318 työhön kelpaamatonta makromolekyyliä. Lopulliseen aineistoon jäi 1847 proteiinia. Proteiiniperheen sisällä sekvenssien identtisyysrajaksi asetettiin 65 %.

Vaatimukset täyttäviä PPII-rakenteita löydettiin 46.6 %:sta jäljelle jääneistä proteiineista. Näistä saatiin vajaat 9000 opetustapausta ja keskimäärin 1500 testitapausta. PPII-rakennetta löydettiin keskimäärin 1.26 % proteiinin pituudelta.

Ongelmana on ennustaa 3-ulotteista rakennetta 2- ja 1 -ulotteisen tiedon perusteella. Tällöin neuroverkon löytymistarkkuudella tarkoitetaan tietyn luokan oikeinennustettujen rakenteiden lukumäärän suhdetta kaikkiin aineistossa esiintyvien tämän luokan rakenteiden lukumäärään. Neuroverkot kykenevät löytämään PPII-luokan tapauksista 72.6 % ja muita rakenteita edustavasta luokasta 74.7 %.

Tietyn rakenteen ennustustarkkuudella tarkoitetaan ennustuksessa oikeinennustettujen rakenteiden lukumäärän suhdetta kaikkiin verkon rakenteiksi väittämien tapausten lukumäärään. Kun neuroverkkoa testataan testijoukolla, jossa on yhtä paljon kummankin luokan tapauksia, saadaan ennustustarkkuudeksi samaa suuruusluokkaa olevat luvut kuin löytymistarkkuudessakin.

Kun ennustetaan luonnollisesti jakautunutta testijoukkoa, on verkko vaikeamman tehtävän edessä. PPII-rakenteiden vähäisen esiintymisen takia etsiytyy PPII-tapausten kaltaisia vastakkaisen luokan alkioita häiritsevästi ennusteisiin. Tämä laskee PPII-rakenteen ennustustarkkuutta. Samasta syystä vastakkaiseen luokkaan voi virheellisesti ennustautua vain vähän PPII-rakenteita. Tämä puolestaan nostaa vastakkaisen luokan ennustetulosta. Kokonaisennustustarkkuus on tasanjakautuneen testijoukon kanssa samansuu-

ruinen. Jos PPII-rakenteen ennustustarkkuutta verrataan esiintymistiheyteen, on ennustustulos parempi kuin esimerkiksi α-kierteellä.

Neuroverkko ennusti PPII-luokkaan runsaasti myös rakenteita, jotka olivat hieman epäsäännöllisiä tai eivät sijainneet tarkalleen ennustuksen määräämässä paikassa. Näitä ei voi PPII-rakenteen määritelmän mukaan pitää oikeina ennusteina. Nämä aineiston ongelmat alentavat ennustetarkkuutta.

Käytännössä PPII-rakenteen ennustaminen on vaikeaa, koska vastakkaisen luokan rakenteita ei pystytä erottelemaan aineistosta tarkasti. Neuroverkkojen ennustustulosta pystytään parantamaan sillä ehdolla, että yhä suurempi osuus kiinnostavista rakenteista jää löytymättä.

Proteiinien 3-ulotteisen rakenteen ja näiden sekvenssien välinen suhde on mielenkiintoinen ja monimutkaisuudessaan ihailtava. Ongelman käsittely antoi runsaasti tietoa tästä biokemian tärkeästä tutkimusalasta ja neuroverkkojen käyttäytymisestä todella vaiken ongelman yhteydessä.

Sisältö

1 Johdanto

Kokeellisesti tapahtuva proteiinirakenteiden selvittäminen on hidasta ja vaatii paljon resursseja. Tämän takia kuilu tuntemattomien ja tunnettujen proteiinirakenteiden lukumäärän välillä on voimakkaassa kasvussa. [23] Raskaan kokeellisen työn takia riittävän hyviä ennustusmenetelmiä etsitään jatkuvasti. Rakenteen ennustaminen on sangen vaikeaa, vaikka ennustusmenetelmät ovat kokeneet jo kolmen sukupolven kehityskaaren. Tilastolliset mallit ovat vaihtuneet älykkyyttä matkiviin menetelmiin.

Proteiinit ovat muodostuneet aminohappoketjuista. Ketjussa on erotettavissa useita rakenteellisia tasoja. Primaarirakenteella tarkoitetaan aminohappojen järjestystä. Järjestystä kutsutaan myös yleisesti sekvenssiksi. Sekundaarirakenteella tarkoitetaan ketjun taipumista säännöllisiin muotoihin. Näitä ovat esimerkiksi α-kierre (α-helix) ja β-säie (β-sheet). Tertiaarirakenteessa proteiinien säännölliset muodot taipuvat toistensa suhteen ja kvartaarirakenteessa proteiinin eri osaketjut liittyvät toisiinsa. [19]

Neuroverkot ovat inhimillistä älyä matkivia järjestelmiä, joissa on monen tieteenalan tunnusmerkkejä. Neuroverkot ovat saaneet alkunsa aivojen teoreettisista malleista. Jo ennen 1900-luvun puoltaväliä aivotutkijat McCulloch ja Pitts kehittivät yksinkertaisia matemaattisia malleja aivojen perusprosessointiyksiköistä neuroneista. Myöhemmin Frank Rosenblatt yhdisti keinotekoisia neuroneita verkoiksi, joita hän nimitti perceptroneiksi. Nämä mallit eivät kuitenkaan pystyneet ratkaisemaan kovin monimutkaisia ongelmia. Rumelhard, McClelland ja Williams saivat alan jälleen uuteen kehitykseen vuonna 1986 esittelemällä takaisinlevityssäännön. Sääntö perustuu epälineaarisen funktion käyttöön keinotekoisissa neuroneissa. [11]

Bioinformatiikan sovelluskentällä reagoitiin varsin nopeasti näihin neuroverkkotutkimuksen saavutuksiin. Rosenberg ja Sejnowski kehittivät menetelmän puheentuottamisen ongelmaan. Työn tuloksena syntyi paljon julkisuutta saanut NETtalk-neuroverkko, joka oppi ääntämään englanninkieltä. Qian ja Sejnowski antoivat alkusysäyksen neuroverkkojen käytöstä sekundaarirakenteiden ennustamiseen, kun he siirsivät vuonna 1988 NETtalk-sovelluksen uuteen ongelmakenttään. Tässä yhteydessä neuroverkolla käsiteltiin englanninkielisten sanojen sijasta proteiinien sekvenssejä. Tuloksena verkko

antoi ennusteen proteiineissa esiintyvistä sekundaarirakenteista. [13] Qian ja Sejnowski ennustivat samanaikaisesti useita rakenteita. Näitä olivat α-kierre, β-säie ja epäsäännölliset rakenteet (coil).

Ennustuksessa pyritään arvaamaan neuroverkolle annetun sekvenssin keskimmäisen aminohapon kohdalla oleva sekundaarirakenne. Neuroverkon ennustustarkkuudella tarkoitetaan oikein menneiden arvausten suhdetta kaikkiin arvauksiin. NETtalk-tyylisellä neuroverkolla tutkijat pääsivät 64,3 %:n ennustustarkkuuteen [13].

Ennustustarkkuutta voidaan parantaa esimerkiksi käyttämällä lisäinformaatiota primaarirakenteen ohella. Tällöin voidaan esimerkiksi hyödyntää tietoa proteiinin kehityshistoriasta. Menetelmässä käytetään hyväksi rakenteiltaan tunnettuja proteiineja, joilla on lähes samat evoluution vaiheet [23]. Lisäinformaationa voidaan käyttää myös tertiaarirakennetta. Tällä tavalla ollaan päästy 79 %:n tarkkuuteen α-rakenteille ja 70 %:n tarkkuuteen β-rakenteille [24].

Tässä työssä tutustutaan polyproliini II-sekundaarirakenteen (luetaan: polyproliini kaksi) ennustamiseen. Polyproliini II eli lyhyemmin PPII on eräs säännöllinen sekundaarirakenne. Azhubei ja Sternberk ovat osoittaneet artikkelissaan rakenteen olemassaolon [1]. Tutkijat toivovat, että PPII-rakenne otettaisiin mukaan myös sekundaarirakenne-ennustuksiin. Artikkeli luo myös edellytykset PPII-rakenteen paikantamiseen DSSP-tiedostosta (Definition of Secondary Structure of Proteins).

Polyproliini II-rakenteen erityispiirteenä on sen selkärangan (backbone) avaruudessa etenevä kolmiomainen rakenne (katso kuva 6). PPII-rakenteen uskotaan osallistuvan solun sisäiseen signaalinvälitykseen ja sillä arvellaan olevan yhteyttä immuunipuutosten syntymiseen [30]. Rakennetta ei esiinny kaikissa proteiineissa. Niissä, joissa sitä esiintyy, on sen esiintymistiheys pieni ja rakenteet ovat lyhyitä. [1] Harvinaisten sekundaarirakenteiden ennustusyrityksiä ei ole tieteellisessä kirjallisuudessa aikaisemmin esitelty.

Tutkielman luvussa kaksi tutustutaan neuroverkkoihin ja neuroverkkojen käyttöön sekundaarirakenne-ennustuksessa. Käytetyt neuroverkot ovat Matlab-laskentaympäristön valmiita funktioita, joten niitä ei tarvitse toteuttaa erikseen. Luvun kaksi lopussa esitellään Matlabin neuroverkkomahdollisuudet sekä tyypillinen opetustapahtuma [3].

Kolmannessa luvussa tutustutaan sovellusalaan. Aluksi tarkastellaan solua, jonka kautta siirrytään proteiinirakenteisiin. Lopuksi esitellään PPII-rakenne.

Laadukas opetusaineisto on neuroverkon käytön kannalta yksi tärkeimpiä asioita. Aineiston hankintaa selvitetään luvussa neljä; lähtökohtana on Protein Data Bank -tietokanta (PDB). Luvussa pohditaan kelpaamattoman

materiaalin poistamista PDB:n tietokannasta ja sekundaarirakenteen paikantamista rakennetiedostoista. Luvun lopuksi pohditaan ainaiston laatua opittavuuden näkökulmasta.

Luvussa viisi esitetään tulokset kahdessa osassa. Aluksi testijoukon luokat ovat yhtäsuuria, jolloin saadaan helppotajuinen tulos neuroverkon ennustustarkkuudesta. Toisessa osassa neuroverkkoa käytetään ennustamaan luonnollisessa jakaumassa olevaa testijoukkoa. Tuloksista syntyneitä johtopäätöksiä esitellään luvussa kuusi.

Työn tavoitteena on selvittää, voidaanko aminohapposekvensseillä opetettuja neuroverkkoja käyttää polyproliini II-sekundaarirakenteen ennustamiseen. Työssä pyritään kiinnittämään huomiota myös ongelmiin, jotka yleisesti tulevat vaivaamaan tällä tavoin tehtävää harvinaisten rakenteiden ennustamista.

2 Neuroverkot työvälineenä

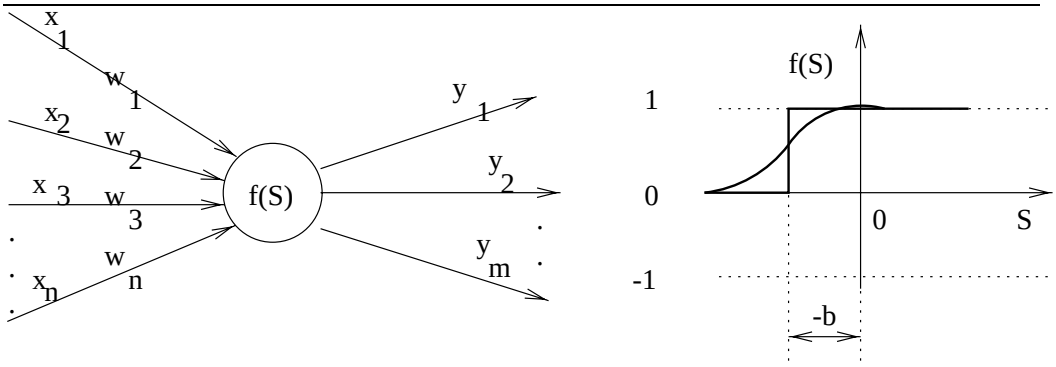
Luku jakautuu viiteen pääteemaan. Alussa tutustutaan neuroverkkojen mallina olevaan aivojen toimintaprosessiin ja rakenteeseen. Myöhemmin siirrytään lähemmäksi keinotekoista prosessointimallia ja esitetään neuroverkon tärkeät elementit sekä yleisimmät sovellusalueet. Luvun loppupuolella tutustutaan neuroverkkojen käyttöön sekundaarirakenne-ennustuksessa. Viimeiseksi esitellään Matlab-ohjelmiston neuroverkkomahdollisuudet.

2.1 Aivot laskennan mallina

Ihminen on kautta aikojen pitänyt älykkyyttään keinotekoisen älykkyyden ihanteena. Toisaalta ihminen on aina verrannut aivojaan monimutkaisimpaan sen hetkiseen tieteen saavutukseen [21]. Satoja vuosia sitten aivoja verrattiin mekaaniseen koneeseen. Renesanssin aikoina aivoja pidettiin yhtä monimutkaisina kuin kellokoneisto. 1800-luvun lopussa aivot olivat ihmisten käsityksissä yhtä monimutkaiset kuin puhelinverkko. Nykyään aivojen monimutkaisuutta on helppo rinnastaa tietokoneeseen - aivotutkimuksen tuloksista tiedetään niiden olevan jotain vielä paljon monimutkaisempaa.

Aivojen rakenteen perusyksikköinä ovat *neuronit*, jotka muodostavat monimutkaisen verkoston. Yksittäiseen neuroniin saapuu yksi tai useita yhteyksiä eli synapseja, joista sähköiset impulssit tulevat syötteinä neuroniin. Neuronin tehtävänä on tehdä päätös syötteiden perusteella. Päätös lähetetään tuloskäytävää eli akonia pitkin jatkokäsittelyyn. Oppiminen ymmärretään neuroneiden välisten yhteyksien syntymisellä, poistumisella ja säätämisellä. Toisin kuin perinteinen tietokone, (aivojen) neuronit pystyvät toimimaan rinnakkaisesti. Tämän takia esimerkiksi kuuleminen ja näkeminen voidaan käsitellä samanaikaisesti. [21]

Vaikka aivot pystyvät käsittelemään rinnakkaista informaatiota menestyksekkäästi, on ongelmia, joita ihmisen on vaikea rinnakkaistaa - esimerkiksi laskutoimituksia [11]. Huomioitavaa on myös, että aivojen toimintaan liittyy inhimillinen tekijä. Aivot eivät toimi algoritmien mukaan, vaan erehtyvät ja oppivat virheistään. Keinotekoiset neuroverkot pystyvät matkimaan tätä erehtyvää ja hiljalleen oppivaa järjestelmää. Valitettavasti neuroverkot ovat



Kuva 1: Keinotekoisen neuroverkon solmu ja bias-skalaarin vaikutus. Solmun vasemmalla puolella ovat syötekanavat ja oikealla puolella tuloskanavat. Solmussa tehdään päätös syötekanavista tulevien syötteiden perusteella ja tulostetaan päätös tuloskanaviin. Bias-skalaari (b) siirtää aktivoitumispistettä.

hyvin spesiaaliin ongelmaan soveltuvia; aivojen kaltaista valtavaa informaatiota sisältävää verkkoa ei ole pystytty rakentamaan.

2.2 Neuroverkon tärkeät elementit

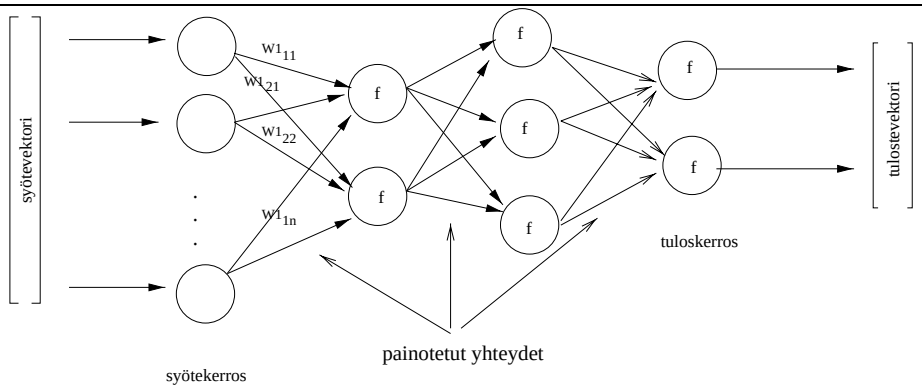
Aivotutkimuksen tarjoamat tulokset ovat olleet lähtökohtana neuroverkkotutkimukselle. Nykyaikaiset neuroverkot eivät muistuta enää alkuperäistä malliansa, vaan yhteys biologiseen esikuvaan on varsin kaukainen. Ideat ovat kehittyneet esikuvasta riippumattomasti. [11]

Neuroverkoissa on erotettavissa kolme pääelementtiä: laskentayksiköt, topologiat ja oppimisalgoritmit [21]. Laskentayksiköt vastaavat aivojen neuroneita ja siksi näitä laskentayksiköitä kutsutaan (keinotekoisiksi) neuroneiksi. Verkon rakennetta kutsutaan topologiaksi tai arkkitehtuuriksi. Rakenteen osalta neuroverkot poikkeavat suuresti aivojen luonnollisista neuroverkoista. Oppimisalgoritmit puolestaan ovat toimintaohjeita oppimiseen.

2.2.1 Neuroni

Yksittäisen neuronin perustehtävänä on laskea arvo S (katso kuva 1). Arvo S muodostuu, kun summataan syötteiden x_i ja painoarvojen w_i tulot. Solmu tuottaa tulosteena summan avulla lasketun suureen. Suure tulostetaan tuloskanavaan tai tuloskanaviin y_i. [11]

Solmun käyttäytyminen on suuresti riippuvainen aktivaatiofunktion f luonteesta. Aktivaatiofunktio voi olla kynnysfunktio, jolloin on mielekästä ajatella solmun olevan “aktivoitunut” tai “aktivoitumaton”. Aktivaatiofunktio voi olla myös lineaarinen tai epälineaarinen esim. sigmoidifunktio. Nämä



Kuva 2: Kolmikerroksinen perceptron-verkko. Vasemmalta verkko saa vektorimuodossa olevan syötteen, joka välittyy syötekerroksen läpi seuraavalle kerrokselle painotettujen yhteyksien kautta. Seuraavissa kerroksissa perceptron-yksiköt laskevat syötteet yhteen ja soveltavat tähän aktivaatiofunktiota f. Uudet impulssit siirtyvät painotettujen yhteyksien kautta seuraavaan kerrokseen. Oikealta saadaan verkon tuloste syötevektorin tapaukseen. Kerrosten välisiä yhteyksiä on helppo käsitellä matriiseina.

funktiot mahdollistavat neuroverkon hienostuneemman oppimisen, jossa käytetään hyväksi funktion differentioituvuutta. Tällöin solmun tulos voi olla myös jokin arvo aktivoituneen ja aktivoitumattoman solmun väliltä. Solmun tasapainoa säädellään bias-skalaarilla. Tämä skalaari siirtää solmun aktivoitumispistettä (katso kuva 1).

Amerikkalainen psykologi Frank Rosenblatt käytti neuroverkon solmuista nimitystä perceptron. Alussa perceptron käsitti vain sisääntulevat painoarvoilla varustetut syöte- ja tuloskanavat sekä kynnystävän funktion. Myöhemmin tästä neuronin perusmallista on alettu käyttää nimitystä klassinen perceptron ja uudemmista malleista käytetään pelkästään nimitystä perceptron. [21]

2.2.2 Topologia

Neuroverkon topologialla tarkoitetaan sen rakennetta eli solmujen lukumäärää, yhteyksien laatua ja eri merkityksessä olevia kerroksia. Tässä yhteydessä käsitellään pelkästään monikerroksisen eteenpäin syöttävän perceptron-verkon rakennetta.

Yksittäinen neuroverkon kerros muodostuu perceptron-laskentayksiköistä. Kerroksien välisestä tiedonsiirrosta vastaavat painotetut yhteydet (katso kuva 2). Eteenpäinsyöttävyys tarkoittaa, että neuroverkossa inpulssin suunta on aina sama. Verkon toisessa päässä on syötepuoli (input) ja toisessa päässä

on tulostepuoli (output). Syötevektorit ohjataan syötepuolelle, josta syöte välittyy prosessoinnin suorittuessa kohti tulospuolta. Tulospuolelta saadaan neuroverkon antama tulos, luokittelu tai ennuste syötevektorille.

Seuraavaa tarkastelua varten esitellään käsite hahmoavaruus. Esimerkkinä yksinkertaisesta hahmoavaruudesta on 2-ulotteinen taso. Jokainen 2-ulotteinen vektori voidaan esittää tasolla yhdellä pisteellä. Tason pisteet voidaan kuvitella edustavan joitakin ilmiöitä, joita halutaan luokitella. Rajaamalla tasolta saman luokan pisteryppäitä, jaetaan hahmoavaruutta osiin. Samalla tavoin neuroverkko rajaa hahmoavaruuden osiin.

Kerroksien lukumäärästä riippuu, kuinka monimutkaisen luokittelun verkko voi tehdä. Verkko, jossa on vain syöte ja tuloskerros, voi muodostaa hahmoavaruuteen konvekseja alueita. Konveksi alue on sellainen, jossa jokaisen pisteen välinen suora pysyy alueen sisällä. Solmujen lisääminen monimutkaistaa alueita, mutta ne pysyvät aina konvekseina. Verkko, jossa edellisen kerroksen lisäksi on yksi piilokerros, muodostaa piilokerroksella konveksit joukot, jotka saadaan tuloskerroksella syötteinä. Nyt tuloskerros voi muodostaa konveksien joukkojen kombinaatioita. Nämä ovat monimutkaisempia kuin konveksit joukot. Kun verkkoon lisätään toinenkin piilokerros, voidaan hahmoavaruuteen muodostaa mielivaltaisia alueita. [11]

2.2.3 Oppiminen

Haykinin mukaan oppimisprosesseista voidaan löytää kaksi pääkategoriaa: algoritmit ja paradigmat. Paradigmat ovat suuren luokan suuntaviivoja oppimiselle. Näistä voidaan erottaa ohjattu-, vahvistava- ja itseorganisoituva-oppiminen. [8] Ohjatussa oppimisessa neuroverkolle annetaan opetusjoukko ja tämän alkioita vastaavat luokitukset. Verkko säätää yhteyksien painokertoimia ja oppii opetusalkioiden piirteitä. Itseohjautuvaa oppimista käytetään, kun syötejoukon luokkia ei tiedetä. Verkon tehtävänä on tunnistaa syötejoukon piirteet. Sekundaarirakenne-ennusteissa luokat yleensä tunnetaan ja oppimisessa käytetään ohjattua paradigmaa.

Oppimisalgoritmit voidaan Haykinin mukaan luokitella seuraaviin tyyppeihin: kilpaileva oppiminen, Hebbin oppiminen, Thorndiken lain mukainen oppiminen, Bolzmannin oppiminen ja virheenkorjaava oppiminen. Monikerroksisessa perceptron-verkossa käytetään virheen korjaavaa oppimista ja virhettä minimoidaan takaisinlevityssäännöllä. Tällöin neuroverkolla on laskentayksiköissä differentioituvat aktivaatiofunktiot. Verkolle annetaan syötevektori ja tämä laskee tuloksen. Tuloksena tullutta arvoa verrataan siihen tulokseen, mikä sen pitäisi olla, ja lasketaan näiden erotus. Erotusta ja aktivaatiofunktioita käyttäen säädetään neuroverkon painokertoimia siten, että

virhe pienenee.

2.3 Neuroverkolle soveltuvat ongelmat

Neuroverkkoja voidaan soveltaa alueille, joissa luokitellaan tapauksia, ennustetaan ilmiöiden todennäköisyyksiä tai segmentoidaan tietoa [11]. Ennustaminen voidaan jakaa kahteen osa-alueeseen: ilmiön olemassaolon ennustamiseen ja aikasarjaennustamiseen. Näistä kaikki muut paitsi segmentointi kuuluvat ohjatun oppimisen kategoriaan.

Luokitustehtävissä neuroverkolle esitetään eri luokkien yksilöitä. Neuroverkon tehtävänä on oppia kuvaus syötteiden joukosta luokkien joukkoon. Opetuksen jälkeen neuroverkon tulisi pystyä luokittelemaan opetusjoukon alkiot ja myös ne alkiot, jotka ovat opetusyksilöiden kaltaisia. Tällöin verkko on oppinut yleistämään. Esimerkkinä tästä on hahmontunnistus.

Ilmiöiden ennustaminen voidaan ajatella luokittelun laajennuksena. Opetusvaiheessa neuroverkolle esitetään joukko tapauksia ja näiden luokat. Luokat ovat neuroverkolle samassa asemassa kuin luokittelutehtävässäkin - neuroverkon tehtävänä on oppia tämä kuvaus.

Ennustustehtävä on kuitenkin suuremman haasteen edessä kuin luokittelutehtävä. Kuvaus opetusalkioiden ja ilmiön välillä on puutteellinen ja käytetty selittäjien joukko ei pysty täysin selittämään ilmiötä.

Kuvaus on puutteellinen esimerkiksi silloin, jos neuroverkolla ennustetaan ihmisen syntymäpäivän ja silmien värin mukaan aikuisiän pituutta. Tietyn kuukauden 24. päivä on luultavasti syntynyt sekä pitkiä että lyhyitä ruskeasilmäisiä lapsia. Pienellä lisäinformaatiolla voitaisiin huomattavasti parantaa ennustusta; lisäinformaatio voisi olla esimerkiksi vanhempien pituudet.

Sekundaarirakenne-ennustuksesta on havaittavissa nämä ennustustehtävän ongelmalliset piirteet. Uusimmissa menetelmissä proteiinin evoluutioinformaatiolla ja tertiaarirakenteella parannetaan kuvausta ja ennustetuloskin paranee.

Segmentoinnissa verkko pyrkii etsimään kohdejoukosta oleellisimmat piirteet ja lajittelee alkioita joukkoihin. Menetelmää voidaan käyttää, kun luokkia ei tiedetä. Toisaalta voi olla, että luokat tiedetään, mutta halutaan tietää luokkia erottelevat piirteet.

2.4 Neuroverkot sekundaarirakenne-ennustuksessa

Proteiinien 3-ulotteinen rakenne määräytyy aminohappojen järjestyksestä [7]. Sekundaarirakenteiden uskotaan määräytyvän kontekstiriippuvaisesti. Täl-

löin rakenne ei ole kiinni pelkästään juuri tarkastelun kohteena olevasta aminohaposta, vaan tästä edelliset ja seuraavat aminohapot ovat vuorovaikutuksessa toistensa kanssa.

2.4.1 Historia

Aikojen kuluessa sekundaarirakenne-ennustuksessa ollaan siirrytty tilastollisista menetelmistä lähemmäksi älykkyyttä matkivia järjestelmiä. Ensimmäiset menetelmät olivat yhden aminohapon menetelmiä. Tilastollisesti selvitettiin, mitä aminohappoja esiintyi minkäkin rakenteen kohdalla. Tästä pyrittiin tekemään ennusteita rakenteen osalta tuntemattomista proteiinisekvensseistä. Näitä menetelmiä kutsutaan ensimmäisen sukupolven ennustusmenetelmiksi. Toisen sukupolven menetelmissä tarkastellaan 11 - 21 vierekkäistä aminohappoa. Aminohapoista laaditaan tilastollinen arvio siitä, kuinka todennäköisesti tietty sekundaarirakenne esiintyy keskimmäisen aminohapon kohdalla. Ennustustarkkuus jäi näillä menetelmillä alle 70 %. [23]

Neuroverkkojen käyttö sekundaarirakenne-ennustuksissa sai alkunsa siitä, että Qian ja Sejnowski näkivät yhteisiä piirteitä koneellisella puheentuottamisella ja sekundaarirakenteen ennustamisella [13]. Puheentuottamisen ilmiö on verrattavissa sekundaarirakenne-ennustamiseen: kirjoitettu teksti muutetaan puheeksi ja lausuminen on riippuvainen suuremmasta kontekstista. Ennustaminen on mielekäs termi, sillä lausuminen riippuu myös asiayhteydestä. Yhteys sekundaarirakenne-ennustamiseen saavutetaan ajattelemalla aminohapposekvenssi kirjoitettuna tekstinä ja sekundaarirakenteet lausumissääntöinä.

2.4.2 Käytetyt arkkitehtuurit

Rakenteiden ennustamisessa on luontevaa käyttää ohjattua oppimisparadigmaa. Toisaalta on myös mielenkiintoista nähdä, millaisia luokkia neuroverkko rakentaa saadessaan itse muodostaa luokittelun.

Ylivoimaisesti eniten sekundaarirakenne-ennustuksissa on käytetty monikerroksisia perceptron-verkkoja. Esimerkiksi edellä mainittu puhetta tuottava NETtalk-verkko oli monikerroksinen perceptron [11]; samoin oli myös Snellin tekemä muunneltu versio sekundaarirakenne-ennustukseen [13]. Samaa arkkitehtuuria ovat käyttäneet myös ryhmä Sacile, Ruggiero, Rauch ja ryhmä Petersen, Bohr, Brunak, Fredholm, Latrup sekä ryhmä Faricelli ja Casadio [4], [20], [24]. Eksoottisempiakin arkkitehtuureja on käytetty, sillä Hanke ja Raich ovat käyttäneet Kohosen itseorganisoituvaa karttaa proteiinien primaarirakenteen kartoitukseen [6]. Kartan organisoitumisen jälkeen primaari-

rakennekartasta paikannetaan sekundaarirakenneryppäät ja kuvaus sekvenssien ja sekundaarirakenteiden välillä on siirtynyt kartan painokertoimiin.

2.4.3 Koodaus

Neuroverkolle tulevan tiedon koodauksella tarkoitetaan symbolisten muuttujien muuntamista numeeriseen muotoon [11]. Luonnossa esiintyvien aminohappojen lukumäärä on 20, joten neuroverkolle täytyy olla 20 eri numeerisessa muodossa olevaa muuttujaa. Koska sekundaarirakennetta kannattaa tarkastella useamman läheisen aminohapon perusteella, syötetään neuroverkolle usea aminohappo kerrallaan. Aminohapot kerätään järjestyksessä tarkasteluikkunan avulla (katso kuva 13 luvussa 4.3.3). Tarkasteluikkunan pituus pidetään koko sekvenssien läpikäynnin aikana samana. Jokaisessa positiossa tarkastetaan, mikä rakenne tarkasteluikkunan keskikohdan alueella esiintyy, ja sekvenssin luokitus määräytyy tämän mukaan. Tarkasteluikkunaa liikutetaan pitkin aminohapposekvenssiä ja aina yhdestä positiosta saadaan yksi opetusjoukon yksilö.

Lähes kaikissa työn lähdeteksteissä on käytetty bittivektorikoodausta. Tällöin aminohaposta muodostetaan bittivektori, jossa on yksi ainoa ykkönen. Ykkösen sijainti osoittaa, mistä aminohaposta on kyse. Näin menettelemällä aminohapoille ei määräydy mitään numeerista suhdetta toisiinsa. Näin pitää ollakin, sillä missään tapauksessa mitään keinotekoista, neuroverkkoa harhauttavaa suhdetta ei saa syntyä.

Koska yhdessä opetusyksilössä on useampi aminohappo, asetetaan bittivektorit toinen toisensa perään. Tällöin esimerkiksi 13-mittaisen tarkasteluikkunan tapauksessa yhdestä bittivektorista tulee 260-ulotteinen bittivektori.

Neuroverkon syötekerrokselle tulee neuroneita $20\times$ tarkasteluikkunan pituus. Syötekerrokselta on mentävä neuroverkon seuraavan tason jokaiselle solmulle painotettu yhteys. Ensimmäiseltä kerrokselta ensimmäiselle piilokerrokselle syntyviä yhteyksiä on 13-mittaisella tarkasteluikkunalla $260\times$ piilosolmujen lukumäärä.

Tästä huomataan, että yhteyksien lukumäärä kasvaa voimakkaasti sekvenssin pituuden ja piilosolmujen lukumäärän kasvaessa. Tämä on tulkittava rajoittavaksi tekijäksi, sillä yhteydet määräävät opetusjoukon kokoa. Opetusjoukon tapausten määrän pitäisi olla kymmenen kertaa neuroverkon yhteyksien lukumäärä [11]. Tällöin aineiston vähyydestä tulee huomattava ongelma. Tämä vaikutti tämänkin työn ratkaisuihin mm. tarkasteluikkunan pituuteeen ja piilosolmujen määrään.

Hanke ja Reich esittävät artikkelissaan fraktaalikoodauksen [6]. Koodauk-

sessa käytetään hyväksi aminohappojen vesisietoisuutta (hydrofiilisyys), jonka avulla aminohapoista muodostetaan 2-ulotteinen ykkösiä ja nollia sisältävä matriisi. Matriisissa luvut 1 sijaitsevat aminohapolle ominaisessa paikassa suhteessa vesisietoisuuteen. Sekvenssin aminohapot muodostavat matriisiin kuvioita, joita neuroverkko oppii tunnistamaan.

2.5 Matlab ja opetustapahtumat

Proteiinimolekyylien karsinnat ja rakenteiden haku tehtiin Linux-ympäristössä. Valmiit opetusryhmän matriisit siirrettiin tämän jälkeen Windows NT-ympäristöön, jossa Matlabilla suoritettiin neuroverkkojen opetus, testaus ja verkon analysointi. Tässä kohdassa tutustutaan hieman tarkemmin Matlab-ohjelmaan ja sen neuroverkkotyökaluihin.

2.5.1 Matlab ja neuroverkot

Matlab-ohjelmisto on MathWorks Inc -yhtiön tuottama kaupallinen matemaattiseen laskentaan ja visualisointiin tarkoitettu laskenta- ja ohjelmointiympäristö. Se on saatavissa useisiin erityyppisiin tietokoneisiin ja käyttöjärjestelmiin mikrotietokoneista keskuskoneisiin. Matlab on saavuttanut suosiota erityisesti tekniikan alalla niin tutkimuksessa kuin opetuksessakin.

Matlab-ohjelmisto on numeeriseen laskentaan rakennettu järjestelmä. Pääasiassa lukujoukot hallitaan matriisien avulla ja laskenta tapahtuu matriisioperaatioita käyttäen. Mukana on runsaasti myös erilaisia valmiita matemaattisia funktioita ja saatavana on monille sovellusaloille omia erikoisohjelmistoja.

Matlab tukee myös neurolaskentaa. Käytössä on runsaasti erilaisia verkkoarkkitehtuureja, oppimisparadigmoja, oppimisalgoritmeja ja visualisointityökaluja.

PPII-ennustuksessa käytettiin eteenpäinsyöttävää kaksikerroksista perceptron-verkkoa. Matlabissa monikerroksinen perceptron muodostetaan peräkkäisillä painomatriiseilla ja *bias*-vektoreilla. Neuroverkon käyttö on tällöin matriisien kerto- ja yhteenlaskua. Perceptron-verkossa on valittavana kerrosten lukumäärä, kerrosten solmujen lukumäärä, bias-vektorin käyttö sekä aktivaatiofunktion luonne.

2.5.2 Omat funktiot

Matlabia voi käyttää komento-ohjattuna, mutta komentoja voidaan kirjoittaa myös komentotiedostoiksi eli funktioiksi. Funktiot ovat kätevä tapa halli-

ta monimutkaisia alustustoimenpiteitä. Työn yhteydessä syntyi joukko funktioita, jotka hallitsevat Matlabin valmiita neurolaskentafunktioita.

Ennen kuin PPII- ja vastakkaisen luokan -tapauksia sisältävät matriisit siirretään NT-ympäristöön, on opetusjoukko jo esiprosessoitu. Aineisto on numeerisessa muodossa, tasamääräinen ja se on vuorottaisessa järjestyksessä. Ensimmäisenä on PPII-tapaus ja seuraavana on vastakkaisen luokan tapaus jne.

Opetusjoukon kohdetapaukset muodostetaan koodaamalla vuoronperään PPII-luokan tunnus ja vastakkaisen luokan tunnus. Luokat esitetään 2-ulotteisina bittivektoreina. PPII-tapausta merkitään vektorilla

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

ja vastakkaisen luokan tapausta vektorilla

$$\begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Opetustapahtuman pääfunktiona käytetään funktiota *opeta_bp*. Tälle annetaan parametreina opetusjoukko, opetusjoukkoa vastaavat luokat, testijoukko ja tämän luokat, rajamatriisi, neuroverkon kerroksissa olevien solmujen lukumäärät sekä opetusiteraatioiden lukumäärä (opetusiteraatio on toimenpide, jossa verkolle esitetään opetusjoukko ja verkko tekee painokertoimiinsa muutoksen). Loput opetustapahtumaa hallitsevat muuttujat ovat valmiina kirjoitettuna *opeta_ bp*-funktiossa. Rajamatriisissa neuroverkolle ilmoitetaan, missä rajoissa muuttujat vaihtelevat. Rajamatriisi on siis korkeudeltaan syötevektorin mittainen ja leveydeltään kahden luvun mittainen.

Funktiossa *opeta_ bp* suoritetaan neuroverkon alustus Matlabin valmiilla $initff$-funktiolla, joka palauttaa alustetut painomatriisit sekä bias-vektorit. Tämän jälkeen alustetaan parametrivektori, joka annetaan opetustapahtumaa hallitsevalle funktiolle parametrina. Parametrivektorissa annetaan opetusvakio, kuvaajan päivityskerrat ja opetuskierrosten lukumäärät. Tämän jälkeen opetus- ja testiaineisto annetaan $trainff\_v$-funktiolle.

Funktio *trainff_ v* on tämän työn yhteydessä kehitetty versio Matlabin omasta funktiosta. Se opettaa neuroverkon alkuperäisen mallin mukaan, mutta käyttää validiointijoukkoa opetustapahtuman tarkkailuun. Tarkkailu tapahtuu seuraamalla neuroverkon tulosteen ja testijoukon virhettä.

Kun verkko on saavuttanut parhaan ennustuskyvyn testijoukon suhteen ja virhe alkaa kasvamaan, opetus keskeytetään. Tällöin *opeta_bp*-funktio palauttaa komentotasolle neuroverkon painomatriisit ja bias-vektorit. Tästä alkaa palautetun neuroverkon suorituskyvyn mittaaminen.

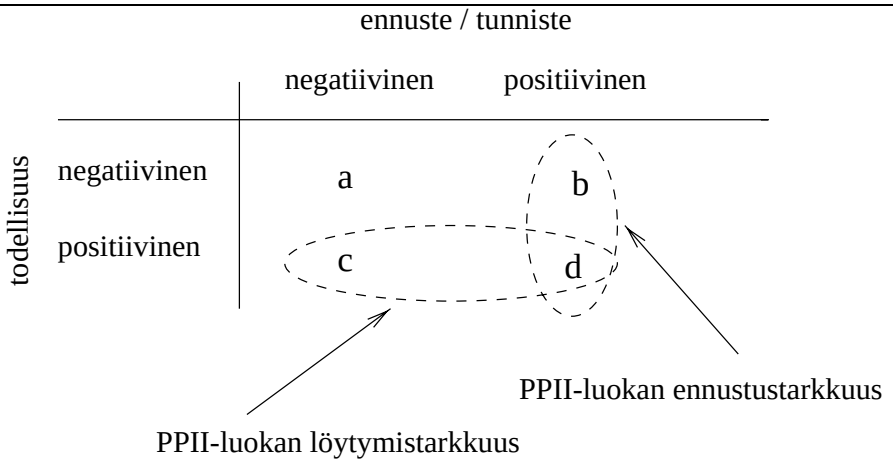
Suorituskyky mitataan *validioi*-funktiolla. Funktio palauttaa lukumäärät kummallekin luokalle tehdyistä oikeista ja vääristä päätöksistä.

Neuroverkon päätös tapahtuu *voittaja saa kaiken* -menetelmällä. Neuroverkko tulostaa syötevektorille arvon kummastakin tulossolmustaan. Luokituksen ratkaisee se, kumman tulossolmun arvo on suurempi.

2.5.3 Tyypillinen opetustapahtuma

Opetustapahtuman kestoon vaikuttavat piilosolmujen lukumäärä ja opetus- ja testiaineistojen koot. Satunnaisen alustuksen vaikutus on myös suuri, sillä samoilla aineistoilla ja solmujen määrillä voidaan joutua tekemään valtavasti toisistaan poikkeavia laskentamääriä. Tästä esimerkkinä erään aineiston opetus, jossa opetusjoukossa oli n. 14000 tapausta. Tällöin neljällä piilosolmulla tehty opetus vaati 460 opetusiteraatiota, kahdeksalla piilosolmulla 3116 iteraatiota ja viidellätoista piilosolmulla 399 iteraatiota. Tapaus kahdeksan joutui erilaisen alustuksen takia tekemään moninkertaisen työn. Keskimäärin opetusiteraatioita tarvittiin noin $1000 - 2000$.

Opetuksen alkuvaiheessa opetusvirhe pienenee varsin nopeasti, mutta hidastuu, kun aineiston pääpiirteet ovat löytyneet. Pääpiirteet löytyvät tavallisesti muutamalla kymmenellä iteraatiolla. Tästä alkaa pitkä ja hidas eteneminen kohti optimaalista ennustuskykyä.



Kuva 3: Neuroverkon ennustetulos. Ennustetulos esitetään neljän luvun kokonaisuutena. Luvut ovat prosentuaalisia osuuksia koko testijoukosta. Symbolilla a merkitään PPII-rakennetta sisältämättömän luokan oikein menneiden ennusteiden prosenttuaalista osuutta. Symbolilla d PPII-luokan oikein menneiden prosenttuaalista osuutta. Symbolit b ja c edustavat vastaavien luokkien väärinennustettujen tapausten prosentuaalista osuutta.

Yhden neuroverkon opetus saatetaan päätökseen tulosten kirjaamisella. Tulokset kirjataan kuvan 3 mukaisesti [14]. Tällöin päädiagonaalilta on helposti luettavissa verkon suorituskyky. Luvut on esitetty prosenttuaalisina osuuksina koko testijoukosta. Verkon ennustuskyky saadaan PPII-luokalle suhteella

$$\frac{d}{b+d} \tag{1}$$

ja vastakkaiselle luokalle suhteella

$$\frac{a}{a+c}. \tag{2}$$

Löytyneiden PPII-rakenteiden osuus voidaan laskea suhteella

$$\frac{d}{c+d} \tag{3}$$

ja vastaavasti vastakkaiselle luokalle

$$\frac{a}{a+b}. \tag{4}$$

3 Millaisia ovat proteiinirakenteet?

Tässä luvussa tutustutaan lyhyesti proteiinien syntyprosessiin ja rakenteisiin. Luvun lopussa esitellään polyproliini II-rakenne.

3.1 Solu

Organismin kudos muodostuu soluista. Eri kudosten ja elinten solut ovat erikoistuneet kukin omaan tehtäväänsä. Solu mielletään usein organismin pienimmäksi toimintayksiköksi. Solu on helppo mieltää myös elävän organismin rakenteisuuden perusyksiköksi. [28]

Solun tumassa sijaitsee informaatio organismin perimästä (DNA). Tuma ja muut solun rakenneyksiköt tarvitsevat ympärilleen soluliman. Solulima on pääosin vettä. [28]

Solun muita tärkeitä osasia ovat tumajyvänen, ribosomit, solulimakalvostot, Golgin laite, mitokondriot jne. Jokaisella näistä on oleellinen merkitys solun toimintaan. Varsinkin ribosomit ovat keskeisessä asemassa proteiinien syntyprosessissa, jossa myös proteiinirakenteet määräytyvät. [28]

Solun koostumusta voidaan tarkastella myös kemiallisten yhdisteiden kautta. Solussa on runsaimmin vettä. Veden tehtävänä on muodostaa ja liuottaa vetysidoksia. Veden lisäksi solu sisältää myös epäorgaanisia aineita, hiilihydraatteja, lipidejä sekä nukleiinihappoja. [29]

3.2 Proteiinisynteesi

Edellisten yhdisteiden lisäksi soluissa on runsaasti proteiineja. Proteiinit toimivat solujen rakenneosina. [29] Proteiinien tehtäviä ovat myös varastointi, kuljetus, viestintä, suojaus ja katalysointi [25]. Ihmiseltä on löydetty noin satatuhatta erilaista proteiinia [28].

Proteiinien muodostuminen (proteiinisynteesi eli translaatio) saa alkunsa solun perinnöllisestä tiedosta eli deoksiribonukleiinihaposta (DNA). DNA sisältää informaation kolmen ribonukleiinihappo-tyypin (RNA) muodostamiseen. DNA ohjaa siis translaatiota. RNA-synteesissä (transkriptio) muodostuu lähetti-RNA, siirtäjä-RNA ja ribosomaalinen RNA. Näistä lähetti-RNA ja siirtäjä-RNA osallistuvat proteiinisynteesiin. [25]

Proteiinisynteesi on monimutkainen tapahtuma. RNA:n lisäksi tapahtumaan osallistuu kymmeniä eri komponentteja. Yksinkertaistaen voidaan ajatella, että translaatiossa siirtäjä-RNA toimittaa aminohapot ribosomille, jossa ne lähetti-RNA:n sanelemassa järjestyksessä liitetään aminohappoketjuksi. Synteesissä ketjulle määräytyy ominainen aminohappojärjestys eli sekvenssi, pituus ja 3-ulotteinen rakenne. Jos yhdessä ketjussa on yli 20 aminohappoa, niin ketjua kutsutaan proteiiniksi; pienempiä ketjuja kutsutaan peptideiksi. [25]

3.3 Aminohapot ja sidokset

Aminohapot ovat toisiaan muistuttavia molekyylejä. α-Hiileen (C_α) on liittyneenä aminoryhmä, karboksyyliryhmä ja sivuketju. Amino- ja karboksyyliryhmät ovat kaikissa aminohapoissa samanlaiset. Poikkeavuuden aiheuttavat ainoastaan sivuketjut, jotka myös siten määräävät aminohapon ominaisuudet. Ominaisuudet liittyvät lähinnä vesisietoisuuteen sekä pH-arvoihin. [25]

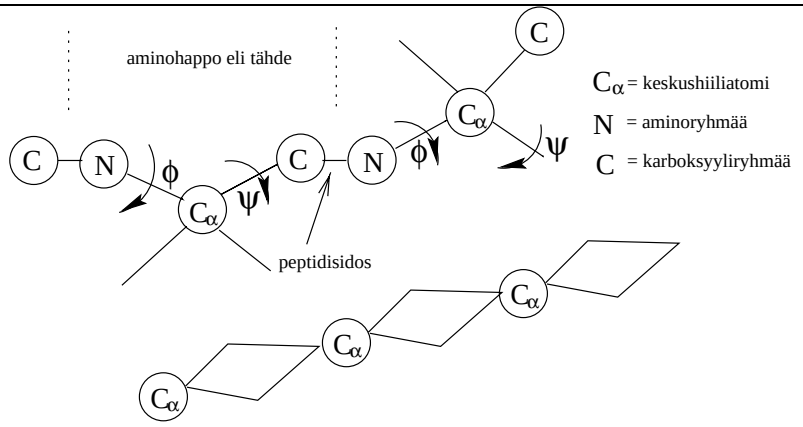
Peptidisidokset syntyvät aminohappojen välille proteiinisynteesissä. Ribosomi liittää edellisen aminohapon karboksiryhmän seuraavan aminohapon aminoryhmään. Reaktiossa syntyy peptidisidos ja yksi vesimolekyyli. Syntyneellä ketjulla on täten aina tietty suunta: alkupäässä on vapaa aminoryhmä ja lopussa vapaa karboksyyliryhmä. [25]

Syntynyt peptidisidos on jäykkä, mutta α-hiilen sidokset pääsevät pyörimään akselinsa ympäri. Tämän takia sivuketjut voivat asettua niille sopivimpaan asentoon ja proteiinimolekyylin 3-ulotteinen rakenne syntyy. [25] Proteiinin rakennetta sekä α- ja β-sekundaarirakenteita on esitelty kuvassa 5.

Aminohapon ympäristöstä käytetään nimitystä tähde. Tähde käsittää keskushiiliatomin ja tämän toisella puolen olevan aminoryhmän sekä toisella puolella olevan karboksyyliryhmän. (katso kuva 4). Ketjun kääntymistä kuvataan edellisen aminohapon ψ- ja seuraavan aminohapon ϕ-kulmalla. [1] Näiden kulmien perusteella voidaan tutkia myös sekundaarirakenteiden muodostumista.

3.4 Proteiinien rakennehierarkia

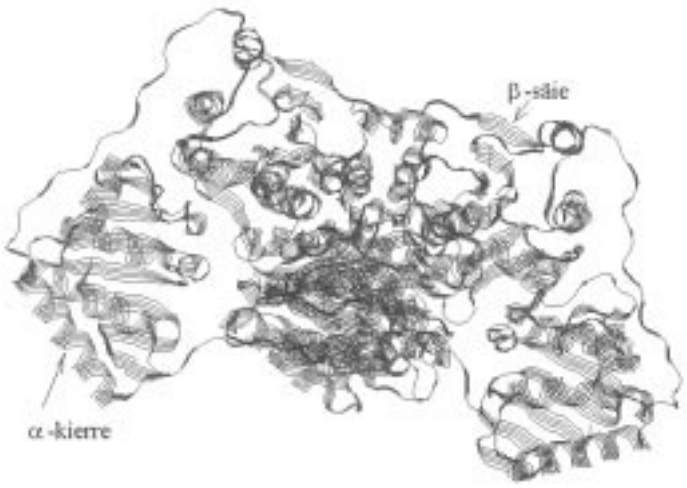
Proteiinien lopullinen rakenne määräytyy neljästä erilaisesta rakenteen tasosta. Primaarirakenteella tarkoitetaan aminohappojärjestystä. Järjestystä kutsutaan yleisesti sekvenssiksi. [25] Proteiinin primaarirakenne voidaan selvittää joko tutkimalla syntynyttä proteiiniketjua tai rakenteen koodaavan geenin perusteella. [28]



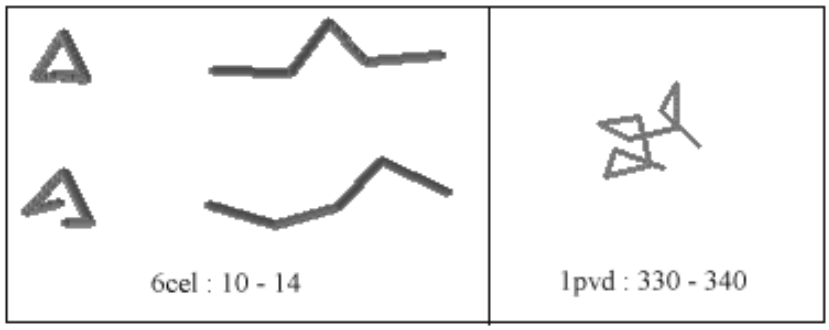
Kuva 4: Polypeptidiketju. Peptidisidos syntyy, kun toisen aminohapon aminopää reagoi toisen aminohapon karboksyylipään kanssa. Peptidisidos on jäykkä, mutta α-hiilen ympärillä olevat kulmat ϕ ja ψ taipuvat. Kuvan alaosassa on havainnollistettu rakenteen tasomaisuutta.

Aminohappoketju muodostaa ketjun sisäisiä ja eri ketjujen välisiä vetysidoksia. α-hiilen sidoksien päästessä vapaasti liikkumaan muodostuu ketjussa paikallisia tunnusomaisia rakenne-elementtejä. Rakenne pyrkii muotoutumaan energian ja rakenteen osalta optimaaliseen muotoonsa [28]. Rakenne-elementistä käytetään nimitystä sekundaarirakenne. Proteiineissa on myös alueita, joissa ei esiinny säännöllisiä sekundaarirakenteita. [25]

Aminohappojen sivuketjujen välille muodostuvat vuorovaikutukset ja ke-



Kuva 5: Atomikoordinaattien perusteella piirretty proteiini 1pvd sekä α- ja β -rakenteiden esittely. Kuva on piirretty RasMol 2.6 ohjelmalla.



Kuva 6: Atomikoordinaateista piirrettyjä PPII-rakenteita. Vasemmalla on lyhyt PPII-rakenne ja oikealla pitkä PPII-rakenne. Kuvat on piirretty RasMol 2.6 ohjelmalla.

mialliset sidokset poimuttavat proteiinia. Poimuttumisen yhteydessä tapahtuvaa taipumista kutsutaan tertiaarirakenteeksi tai luonnolliseksi muodoksi. Kun kaksi tai useampia tertiaarirakenteita liittyy suuremmaksi rakenteeksi, syntyy kvaternaarirakenne. [28]

3.5 PPII on harvinainen sekundaarirakenne

Proteiineissa esiintyy joitakin harvinaisempia säännöllisiä alueita, jotka eivät kuulu α-kierre eikä β-säie-rakenteisiin. Näitä ovat esimerkiksi 3_{10}-kierre, π-kierre ja polyproliini I ja II (PPI ja PPII).

Polyproliini II löydettiin, kun kollageeni-proteiinista löydettiin vasenkätisestä α-kierteestä poikkeava vasenkätisesti taipuva sekundaarirakenne. Tätä rakennetta alettiin kutsua polyproliiniksi, koska runsaimmin rakenteen kohdalla esiintyy proliini-aminohappoja. Teoreettisten laskujen mukaan myös oikealle taipuva polyproliini-rakenne (PPI) on mahdollinen, mutta tätä ei tavallisesti luonnossa tavata. [27] PPII-rakenne voi syntyä, vaikka proliinia ei lähistöllä olisikaan. Azhubei ja Sternberg ovat analyysissään löytäneet 96 rakennetta, joissa 30 %:ssa ei esiinny proliinia [1].

Polyproliini II taipuu vasemmalle edeten avaruudessa kolmiomaisin rakentein (katso kuva 6). Rakenteessa esiintyy kolme aminohappoa kierroksella. Pääasiassa rakenteet ovat lyhyitä; enimmäkseen kolmesta viiteen aminohappoa pitkiä.

Turpeenoja esittelee kirjassaan kollageenissa esiintyvää erikoista rakennetta: ”Rengasrakenteinen proliini ei voi osallistua α-kierteen muodostamiseen, vaan kollageenimolekyylissä kolme polypeptidiketjua asettuvat rinnakkain ja

liittyvät yhteen vetysidoksilla. Kollageenin vetolujuus perustuu juuri tähän kolmen polypeptidiketjun muodostamaan sekundaarirakenteeseen." [28]

Tämä voisi selittää sen, miksi PPII-rakenteella on huomattu olevan muita sekundaarirakenteita hajottava ominaisuus. Tässä myös vahvasti viitataan siihen, että rinnakkaiset polypeptidiketjut osallistuvat sekundaarirakenteen muodostamiseen. Aminohappo ja sen ympäristö eivät riitä selittämään rakenteen muodostumista, vaan rakenteen syntymiseen tarvitaan rinnakkaisia ketjuja.

Polyproliini II:lla on immunologisia- ja viestien välittämiseen liittyviä tehtäviä. Polyproliini II on monessa mielessä kiinnostava rakenne. Kiinnostusta lisää myös se, että rakenne ei ole ollut kohteena aikaisemmissa sekundaarirakenne-ennusteissa. [30]

4 Aineiston hankinta

Luvun alussa tarkastellaan proteiinirakennepankista (Protein Data Bank, PDB) haettujen molekyylien karsintaa. Karsinnassa poistetaan DNA- ja RNA-molekyylit, identtiset ja epätarkat proteiinit. Seuraavaksi esitellään sekvenssien identtisyysvertailua. Menetelmällä karsitaan aineiston sukulaisproteiinit. Luvun keskivaiheilla käsitellään PPII-rakenteita edustavien sekvenssien etsimistä rakennetiedostosta. Tässä vaiheessa otetaan myös käyttöön termi ei-PPII-luokka. Luokassa on muita kuin PPII-rakennetta edustavia sekvenssejä. Nämä ei-PPII-rakenteet poimitaan samanaikaisesti PPII-rakenteiden kanssa. Luvun lopuksi selvitetään aineiston ominaisuuksia.

4.1 Tietokannan perkaus

Protein Data Bank eli PDB-tietokanta sisältää tietoa biologisista makromolekyyleistä - pääasiassa proteiinimolekyyleistä. Tietokannassa on atomikoordinaatteja, rakennetietoa sekä yleistä tietoa tietokannan makromolekyyleistä. Marraskuussa 1998 saatavana oli 8165 makromolekyylin tiedot.

Neuroverkkojen tarvitseman materiaalin pitää olla monipuolista, laadultaan hyvää ja sitä on oltava runsaasti. Aineiston pitää olla monipuolista, koska PPII-rakenteita saattaa esiintyä lukuisissa proteiiniperheissä (samoja ominaisuuksia sisältävien proteiinien joukko) ja monenlaisissa aminohappojaksoissa.

Opetusaineiston laatu on tärkeä ominaisuus neuroverkkojen opetusprosessissa. Huonolaatuinen opetusaineisto saattaa ohjata neuroverkon harhaan. Yksittäisen makromolekyylin tietojen laadukkuus ilmoitetaan resoluutioarvolla.

Tietokannan sisältö luetellaan *entries*-tiedostossa, joka voidaan hakea PDB:n kotisivulta. Tiedostossa esitetään ASCII-muodossa jokaisen tietokantaan liitetyn makromolekyylin identifiointitunnus, proteiiniperhe, päivämäärä, proteiinin nimi, lähdeorganismi, tutkijoiden nimet, resoluutio sekä tutkimusmenetelmä. Yhden molekyylin tiedot ovat tiedostossa yhdellä rivillä. Tiedot on erotettu toisistaan tabulaattori-merkillä.

4.1.1 Perkausohjelma

Aineiston perkaukseen tehty ohjelmisto kehitettiin C++ ohjelmointikielellä Gnu-ohjelmistoperheen G++-implementaatiolla. Ohjelmakoodi jakautui seitsemään eri lähdetiedostoon.

Rakenteellinen ohjelmointi soveltuu tähän asetelmaan hyvin. Makromolekyylin kuvaamiseen muodostetaan luokka ja kokonaisuutta hallitaan tietorakenteiden avulla. Luokalla on attribuutteina makromolekyylin ominaisuudet. *Entries*-tiedosto rakennetaan linkitetyksi listaksi. Tiedoston riviä edustaa aina yksi listan alkio.

Redundanssin runsaus aiheuttaa ongelmia *entries*-tiedoston käsittelyssä. Jokaisen molekyylin kohdalla on kirjoitettu sen nimi ja muut tarkentimet aina uudestaan. Nimet ovat monimutkaisia ja niiden kirjoitusasuista on useita muunnelmia.

Tiedostosta löytyi esimerkiksi merkinnät HYDROLASE(O-GLYCOSYL) ja HYDROLASE (O-GLYCOSYL). Nämä saattavat vaikuttaa varsin samankaltaisilta ja tarkoittavatkin samaa makromolekyyliä. E- ja (-merkkien välissä olevan välilyönnin johdosta ohjelmointikielten merkkijonovertailut tunnistavat edelliset esimerkit eri merkkijonoiksi. Ongelmia aiheuttavat myös tiedostossa ilmoitettujen lähdeorganismien nimet. Joissain tapauksissa organismi ilmoitetaan merkkijonon ORGANISM_SCIENTIFIC: jälkeen, mutta joissain tapauksissa tätä tunnistetta ei esiinny. Saattaa olla, että algoritmit eivät tavoita kaikkia mahdollisia tapauksia.

4.1.2 DNA, RNA ja proteiinikompleksit

PDB sisältää joitakin DNA- ja RNA-molekyylejä. Nämä molekyylit voidaan poistaa aineistosta. Karsinnassa kysytään jokaiselta makromolekyyliä edustavalta oliolta, sisältääkö kyseinen *entries*-tiedoston rivi merkkijonoja DNA, RNA tai RIBONUCLEIC(ACID). Jos ehto täyttyy, alkio poistetaan listasta.

Ohjelma löysi reilusti enemmän ehdot täyttäviä rivejä kuin PDB:n kotisivulla ilmoitetaan. Makromolekyylin perhettä kuvaavasta sarakkeesta löytyi 522 RIBONUCLEIC-merkkijonolla varustettua makromolekyyliä. Tämän lisäksi tiedostossa oli 538 DNA-merkkijonon sisältävää riviä ja 231 RNA-merkkijonon sisältävää riviä. Kaikkiaan RIBONUCLEIC, DNA- ja RNA-merkkijonoja sisältäviä rivejä oli 1291. Jäljellä olevassa tietokannassa esiintyy vielä 4 kappaletta DNA/RNA-merkkijonoja.

Karsintaoperaation seurauksena aineistosta poistui 1295 makromolekyyliä. Osa näistä karsiutui aiheettomasti, sillä jälkeenpäin karsittujen joukosta löytyi molekyylejä, jotka olisivat saaneet jäädä aineistoon. Näistä esimerkkinä

DNA:ta prosessoiva entsyymi (2ADM METHYLTRANSFERASE), joka karsiintui rivillä olevan DNA-merkkijonon takia.

Myöhemmin selvisi, että DNA- ja RNA-molekyylit ovat varsin lyhyitä ja tällöin sekvenssin pituutta olisi voitu käyttää karsintaehtona. Sekvenssin pituutta ei ilmoiteta *entries*-tiedostossa, vaan tieto on saatavissa esimerkiksi FASTA-tiedostossa (katso luku 4.2.1).

Karsinta poistaa myös DNA- ja RNA-molekyylit, jotka ovat kompleksina proteiinin kanssa. Tällöin tiedoston rivillä esiintyy merkkijono COMPLEXED. Nämä siirrettiin ajon aikana omaan tiedostoon. Kompleksina olevat proteiinit tutkittiin tapaus kerrallaan asiantuntijatarkastelussa.

Ajon yhteydessä tiedostoon taltioitui 94 DNA- tai RNA-kompleksia. Näistä 20 täytti kelpoisuusehdon. Nämä siirrettiin takaisin aineistoon. Aineistoon jäi karsinnan jälkeen 6821 makromolekyyliä.

4.1.3 Resoluutio

Resoluutiolla tarkoitetaan sitä, kuinka pieniä yksityiskohtia tutkimusmenetelmällä voidaan havaita. *Entries*-tiedoston kullekin proteiinille osoitettua resoluutio-arvoa käytetään karsintaehtona, kun halutaan varmistaa materiaalin laadukkuus.

Entries-tiedostoa ladattaessa jokaisen rivin resoluutiosarake vaatii erityistä tarkastelua. Jos sarakkeessa on merkkijono NOT, niin kyseessä on teoreettinen malli, joka on saatu tietokonesimulaatiolla. Tällä ei voi olla resoluutioarvoa. Jos sarakkeessa on merkkijono NMR, niin molekyylin analysointi on suoritettu ydinmagneettiresonanssi-menetelmällä. Tälle ei ole myöskään resoluutioarvoa.

Teoreettiset mallit poistetaan, mutta NMR-menetelmällä saaduista molekyyleistä voi jäädä mukaan yksi jokaisesta samannimisestä proteiinista. Karsittaviksi asetettiin myös kaikki, joiden resoluutioarvo ylittää 2,5 Å; siis molekyylin laadukkuus ei ole tarpeeksi korkea.

Koko aineistossa esiintyi 1369 makromolekyyliä, joiden resoluutio oli suurempi kuin 2,5 Å. Näistä 301 kuului edellisessä ryhmässä karsittuihin, joten tässä vaiheessa poistettiin 1068 makromolekyyliä. Resoluutio-karsinnan aikana poistui myös aineiston teoreettiset mallit, joita oli tietokannassa kaikkiaan 192 kpl. Tätä edellinen karsinta poisti malleista jo 25 kappaletta, joten tässä vaiheessa malleja poistettiin 167 kpl. Teoreettisten mallien karsinnan ehtona oli, että resoluution paikalla on merkkijono NOT eikä rivillä esiinny merkkijonoa NMR. Tämä menetelmä antoi saman mallien lukumäärän kuin riviltä suoritettu THEORETICAL-merkkijonon haku. Aineistossa on tämän operaation jälkeen jäljellä 5568 makromolekyyliä.

Algoritmi 1 Saman organismin saman nimisten proteiinien poistaminen

1: vertaa ovatko nimet identtiset
2: jos 1. tosi, ovatko yhdisteet samasta organismista,
3: jos 2. tosi, poistetaan yhdiste, jolla on huonompi resoluutio,
4: ota uusi vertailtava pari ja mene kohtaan 1.

4.1.4 Saman organismin identtiset proteiinit

PDB sisältää identtisiä proteiineja, jotka ovat peräisin samasta organismista. Identtiset tapaukset voidaan poistaa aineistosta. Tämä tehdään algoritmin 1 mukaisesti.

Entries-tiedoston *yhdiste*-sarakkeessa olevat tiedot ovat epäselviä. Jokaisesta tapauksesta ei selviä, missä menee yhdisteen nimen ja mahdollisten tarkentimien välinen ero. Tämä on otettava huomioon jo tiedoston latausvaiheessa. Aluksi neljänneltä sarakkeelta etsitään tunnistetta MOL_ID: 1; MOLECULE:. Jos tunniste löytyy, ladataan tunnisteesta 50 seuraavaa merkkiä olion *nimi*-attribuuttiin ellei tabulaattorimerkkiä esiinny tätä ennen. Jos tunnistetta ei esiinny, alkaa etsitty nimike oletettavasti sarakkeen alusta. Tällöin kopioidaan sarakkeen 50 ensimmäistä merkkiä em. muuttujaan tai kunnes tabulaattorimerkki tulee vastaan.

Samoin menetellään ladattaessa lähdeorganismin nimeä. Tässä tapauksessa etsitään merkkijonoa ORGANISM_SCIENTIFIC:, jonka jälkeen mainitaan organismi, josta näyte on otettu - esimerkiksi HOMO SAPIENS.

Algoritmilla 1 ja em. oletuksin koko tietokannasta löytyi 4009 poistettavaa makromolekyyliä. Kun operaatio kohdistettiin yksistään DNA ja RNA -merkkijonoja sisältäville yksilöille, poistettavaksi tuli 795 molekyyliä. Kun aineistosta oli DNA- ja RNA -sisältöiset sekä suuriresoluutioiset molekyylit poistettu, tuli tämän karsintaoperaation yhteydessä poistettua 2649 kappaletta.

Aineistosta on tähän mennessä poistettu DNA- ja RNA-molekyylit, suurilla resoluutioilla varustetut tapaukset sekä samannimiset molekyylit, jotka ovat lähtöisin samasta organismista. Kaksikymmentä proteiinikompleksia on lisätty ensimmäisen karsintaoperaation yhteydessä, joten jäljellä on 2937 makromolekyyliä.

Tietokannassa on mukana myös muitakin rakenteita kuin DNA-, RNA- ja proteiinirakenteita. Nämä täytyy etsiä manuaalisesti asiantuntijan avulla ja poistaa aineistosta.

Poistettavia löytyi muutamia, kuten esimerkiksi TEXTURE OF CONNECTIVE TISSUE-luokituksen saaneet AGAROSE- ja IOTA CARRAGEE-

NAN-yhdisteet. Nämä ovat kuitenkin poistuneet jo aiemmissa vertailuissa.

4.1.5 Siirtyminen identtisyysvertailuun

Kun ohjelma on tehnyt alustavan karsintaketjun, on aineisto jaoteltava proteiiniperheisiin, joiden kesken identtisyysvertailu suoritetaan. Jaottelussa on uudelleen ongelmana merkintöjen epästandardisuus. Tämän takia jouduttiin rakentamaan uusi menetelmä perheiden luokitteluun.

Tavoitteena oli, että luokat saataisiin mahdollisimman luontevasti erilleen ja kuitenkaan luokituksesta ei tulisi liian herkästi erottelevaa. Jos menetelmä on liian erotteleva, luokkia tulee liian paljon. Ohjelman on huomattava kuitenkin olennaiset erot perheiden välillä. Ohjelman on myös poistettava COMPLEX-merkinnät, luettava sulkujen sisältä ja valittava perhettä kuvaava sana.

Perheitä syntyi 506 kpl. Osa oli yhden proteiinin synnyttämiä, mutta osasta perheistä tuli suhteellisen suuria. Suurimmat perheet olivat HYDROLASE- sekä OXIDOREDUCTASE-perheet. HYDROLASE-perheessä oli yli 400 kpl. ja OXIDOREDUCTASE-perheessä oli n. 200 kpl.

Perheen edustajien identifiointikoodit talletettiin *perheet*-nimiseen tiedostoon, jonne syntyi 506 luokkaa. Jokainen luokka on omalla rivillään, jossa yksittäisen molekyylin identifiointitunnus on eroteltu tabulaattoorimerkinnällä toisistaan.

4.2 Identtisyysvertailu

Jokaiselle perheelle tehdään jäsenten kesken parittainen vertailu. Ohjelma saa syötteenä kaksi proteiinia ja palauttaa parin välisen identtisyyden. Jos identtisyys on liian suuri, poistetaan näistä suuremman resoluution omaava. Muutoin kumpikin saa jäädä aineistoon.

Alunperin tarkoituksena oli käyttää Suomen tieteellisen laskennan palveluja ja erityisesti GCG-ohjelmistoperheen GAP-ohjelmaa. Ongelmaksi muodostui kuitenkin GCG-ohjelmiston tiedostomerkinnät, jotka eivät täsmää PDB:stä saatujen merkintöjen kanssa. PDB käyttää makromolekyylin tiedostonimenä ja tunnisteena identifiointitunnusta, jonka mukaan tämän työn aikaisempi analyysi on tehty. GCG ei tunnista tätä formaattia, joten PDB:n tunnisteita ei voida käyttää. Tämän takia GAP-ohjelma rakennettiin itse.

4.2.1 FASTA-formaatti

Identtisyysvertailuja varten tarvittiin oikeassa formaatissa olevat sekvenssit, joissa aminohapot esitetään yhdellä kirjaimella. Käyttöön otettiin FASTA-formaatti ja formaatin mukainen tiedosto haettiin PDB:n internet-sivulta.

FASTA-tiedostossa proteiinit ovat paikannettavissa molekyylin identifiointitunnuksen avulla. Tämä esitystapa sopi tarkoitukseen mainiosti. Tiedoston rivit ovat standardilevyisiä ja rivin pituudeksi on määrätty 80 merkkiä. Koska tiedostossa ei esitetä proteiinien resoluutioarvoa, täytyy tämä arvo hakea *entries*-tiedostosta.

FASTA-tiedoston lukemiseen kehitettiin uusi ohjelmakomponentti. Komponentti sovitettiin identtisyysvertailua tekevän komponentin sekä aineiston hankinnassa käytetyn komponentin kanssa samaan kokonaisuuteen.

Sekvenssin haku toteutettiin mahdollisimman yksinkertaisesti. Tiedostoa luetaan rivi kerrallaan; jos rivillä esiintyy haettu id-tunnus, niin kopioidaan tunnuksen yhteydessä oleva sekvenssi. Molekyylin etsiminen saattaa hieman kuluttaa aikaa, sillä läpikäytäviä rivejä on noin 40 000. Toinen mahdollisuus olisi ollut ladata koko tiedosto keskusmuistiin ja suorittaa haku siellä. Tämä olisi kuitenkin kuluttanut liikaa identtisyysvertailun tarvitsemaa keskusmuistia.

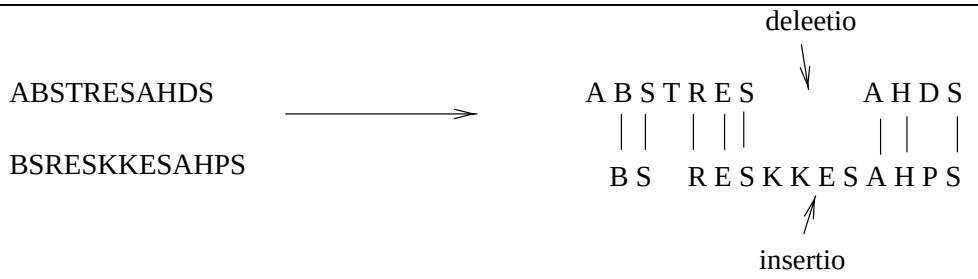
Identtisyysvertailussa muodostetaan viisi matriisia. Matriiseja tarvitaan tietojen taulukointiin ja polun tallettamiseen. Algoritmin kuvauksessa nämä on typistetty kahteen matriisiin.

Jokaisessa matriisissa toinen proteiini edustaa matriisin vaakarivejä ja toinen sarakkeita. Tällöin n. 1000 aminohappoa pitkät sekvenssit tarvitsevat viisi miljoonan solun matriisia.

4.2.2 Kahden sekvenssin rinnastus

Needelman ja Wunch julkaisivat vuonna 1969 sekvenssivertailuja koskevassa artikkelissaan menetelmän, jolla kaksi sekvenssiä voidaan asettaa sukulaisuuden suhteen parhaiten vastakkain [18]. Vastakkainasettamisen voi mieltää kahden merkkijonon asettamista päällekkäin siten, että suurimmat samankaltaisuudet tulevat toistensa kohdalle. Menetelmä voi siirtää tai katkaista sekvenssejä saadakseen parhaiten täsmäävät aminohappoparit vastakkain. Deleetio syntyy sekvenssin sille kohdalle, jossa ei esiinny toisen sekvenssin aminohappoja. Insertio syntyy vastaavasti katkenneen sekvenssin tyhjään tilaan (katso kuva 7).

Menetelmä soveltaa dynaamisen ohjelmoinnin periaatetta, jolla saadaan valtavan suuri laskutoimitusten joukko pienenemään huomattavasti. Menetelmässä toinen proteiinisekvenssi (sekvenssi 1) saa käyttöönsä matriisin sarakkeet ja toinen proteiinisekvenssi (sekvenssi 2) saa rivit. Sarakkeet saaneen proteiinin yksittäinen aminohappo saa käyttöönsä yhden sarakkeen ja vastaavasti toisen proteiinin aminohapot saavat kukin käytettäväkseen yhden



Kuva 7: Sekvenssien rinnastus. Kaksi sekvenssiä asetetaan toisiaan vasten siten, että suurimmat sukulaisuudet tulevat toisiaan vasten. Sekvenssejä voidaan katkaista, jolloin syntyy insertio ja deleetio. Katkeamiskohtien lukumäärää ja pituutta rajoitetaan sakotuksella.

rivin (katso kuva 8). [18] Kun sekvenssit on rinnastettu, voidaan laskea kahden sekvenssin välinen identtisyysarvo.

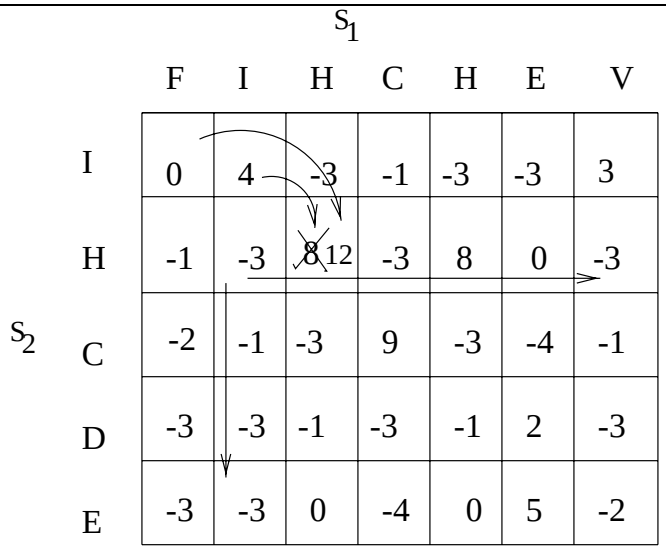
4.2.3 Algoritmin tarkka kuvaus

Tässä esitettävä algoritmi on tehostettu versio Needelmanin ja Wunchin algoritmista. Mukaan on otettu aminohappojen sukulaisuutta kuvaavat luvut, jotka saadaan PAM250-taulukosta. Vastaavaa taulukkoa käytetään myös GCG-ohjelmistossa [17].

Merkitään, että S_1 on sekvenssi 1 ja S_2 on sekvenssi 2 ja, että $S_1(j)$ tarkoittaa S_1:n aminohappoa, joka on järjestyksessä j:s, ja $S_2(i)$ tarkoittaa S_2:n aminohappoa, joka on järjestyksessä i:s. Merkitään vielä, että $M(i, j)$ tarkoittaa taulukon solua, joka on i:s rivi ylhäältä ja j:s sarake vasemmalta. M(i,..j-1) tarkoittaa, että i:nnen rivin sarakkeet käydään 1:stä j-1:een ja M(..i-1,j) tarkoittaa, että j:nnen sarakkeen rivit käydään 1:stä i-1:een. Olkoon sekvenssien pituudet $\|S_1\| = m$ ja $\|S_2\| = n$. Tällöin toiminta etenee algoritmin 2 mukaisesti (katso myös kuva 8).

Aluksi muodostetaan matriisi, jonka korkeus on n ja leveys on m. Matriisi täytetään aminohappojen sukulaisuutta kuvaavilla luvuilla. Samalla voidaan muodostaa myös algoritmin kohdassa 2 tarvittu polkumatriisi P. Matriisin P jokaiseen soluun täytyy pystyä tallettamaan kaksi lukua eli koordinaatit, mistä tähän soluun lisättävä löytyy. Eräs mahdollisuus on käyttää 3-ulotteista matriisia, jonka koko on tällöin $n \times m \times 2$.

Matriisin M jokaiseen soluun muodostuu edeltävän polun kumulatiivinen summa. Kun matriisin alimmainenkin rivi on käyty läpi, on alimmalle riville tai oikeanpuoleisimmalle sarakkeelle syntynyt suurin luku t (tai useita suurimpia lukuja). Tällöin polkumatriisista P nähdään, minkä solujen kautta



Kuva 8: Taulukko sekvenssien rinnastukseen. Vaaka- ja pystyrivillä olevien kirjainten risteyskohtaan merkitään aminohappojen samanlaisuutta kuvaava luku. Samanlaisuusluku saadaan PAM250-taulukosta. Kuvassa on nuolilla esitetty algoritmin suorituksen alkua. Kaarevat nuolet osoittavat vertailuja ja suorat nuolet osoittavat toimituksen etenemistä.

polku on syntynyt. Polkumatriisista nähdään myös, milloin polkuun on syntynyt hyppäys eli sekvenssien rinnastuksessa on syntynyt insertio ja deleetio.

Vaiheessa kolme paikannetaan polun synnyttämä suurin luku. Vaiheessa neljä lasketaan karsintaehdon testausta varten identtisyysarvo. Identtisyysarvo syntyy polulussa olevien identtisten aminohappoparien lukumäärän ja polussa olevien kaikkien aminohappoparien lukumäärän suhteesta.

Kustannus lasketaan sakkotermillä $k = a+bl$, missä a on hypyn avauskustannus, b on hypyn jatkokustannus ja l on hypyn pituus. Hyppy syntyy, jos kahden peräkkäisen position (p_1, p_2) ja (r_1, r_2) vähennyslaskussa jompaan kumpaan positioon $(p_1 - r_1, p_2 - r_2)$ syntyy itseisarvoltaan ykköstä suurempi luku. Toisin sanoen hyppy syntyy, kun kohdan 2 max-funktion yhteenlaskettava saadaan yhden sarakkeen tai rivin yli hypättäessä. Ajoissa käytettiin hypyn aloituskustannuksena lukua 4 ja jatkokustannuksena lukua 2.

Tämä menetelmä on laajennettavissa myös kahta useammalle sekvenssille. Tällöin kolmen sekvenssin samanaikaisessa vertailussa 2-ulotteinen matriisi muutetaan 3-ulotteiseksi ja neljän tapauksessa 4-ulotteiseksi. Menetelmä tulee kyllä laskennallisesti raskaaksi, sillä 2-ulotteisen rinnastuksen yhteydessäkin dynaamisella menetelmällä joudutaan tekemään paljon laskentaa. [5] (Algoritmin kompleksisuustarkastelu on esitetty liitteestä 1).

Algoritmi 2 Sekvenssien rinnastus

1: alustetaan matriisi M PAM-taulukosta saaduilla sukulaisuusluvuilla,
2: käydään järjestyksessä rivit i=2...n; käydään jokaisella rivillä sarakkeet j=2...m ja lasketaan $M(i,j) = M(i,j) + max(M(i-1,..j-1), M(..i-1, j-1))$; jos hyppy ylittää yhden sarakkeen tai rivin, lisätään edelliseen sakkotermi $a + lb$; talleta polkumatriisiin $P(i,j) = (p_1, p_2)$ mistä max löytyy,
3: etsi suurin arvo t alimmalta riviltä tai oikeanpuolimmaiselta sarakkeelta (tähän positioon päätynyt polku on paras),
4: identtisyys $I = \frac{q}{s}$, missä q on polussa olevien identtisten aminohappoparien lukumäärä ja s on polussa olevien solujen lukumäärä.

4.2.4 Ohjelmiston arkkitehtuuri

Tiedonhallintaa organisoivan järjestelmän täytyy hallita myös muita osa-alueita kuin pelkän identtisyyden vertailun. Järjestelmän täytyy lukea kolmea tiedostoa ja täydentää ajon aikana tulostiedostoa. Tulostiedostoon talletetaan vertailtujen proteiiniperheiden alkiot.

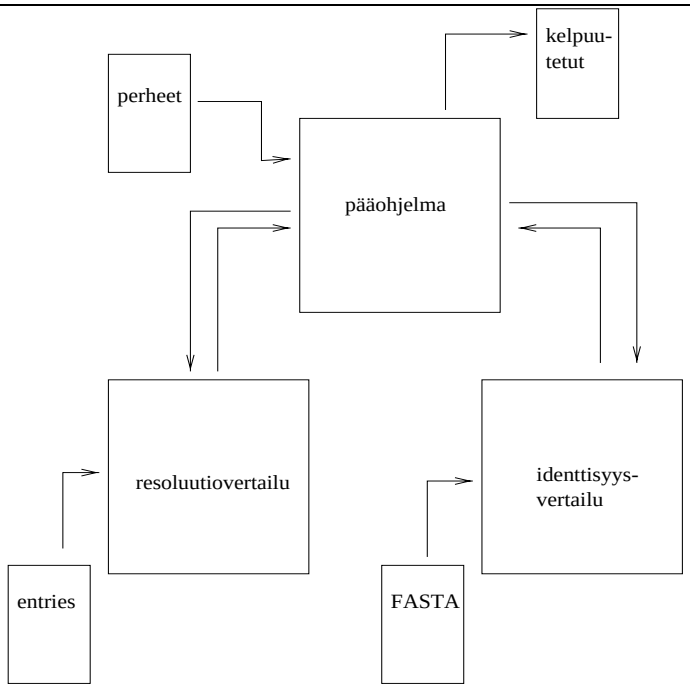
Tiedon organisoinnista vastaava järjestelmä on yhteydessä *perhe*- ja *karsitut*-tiedostoon. Se on yhteydessä myös identtisyyden ja resoluution vertailusta vastaaviin järjestelmiin (katso kuva 9).

Ohjelma hakee järjestyksessä *perhe*-tiedostosta yhden rivin ja lähettää proteiiniparin kerrallaan identtisyysvertailuun. Jos identtisyys ylittää rajan 65 %, komponentti lähettää parin resoluutiovertailuun. Kun parin molempien proteiinien resoluutiot tiedetään, voidaan huonompi poistaa ja resoluutioltaan paremman proteiinin tunnus kirjoitetaan *kelpuutetut*-tiedostoon. Jos identtisyys on alle rajan, kirjoitetaan molempien proteiinien tiedot *kelpuutetut*-tiedostoon.

4.2.5 Identtisyysvertailun tulokset

Tiedoston alkuosassa oli lyhyiden proteiinisekvenssien id-koodeja. Ongelmat paljastuivat, kun vastaan tuli ensimmäinen isompi perhe. Jäseniä saattoi olla useampia kymmeniä ja sekvenssit satoja merkkejä pitkiä. Tällöin yhden sekvenssiparin vertailu saattoi kestää kymmeniä sekunteja. Keskimäärin parittaiset identtisyydet vaihtelivat 30 - 40 %:n välillä, joten karsintaa suoritettiin suhteellisen vähän. Tällöin jo kolmenkymmenen jäsenen kokoisen perheen yhteydessä joutui suorittamaan yli 400 parittaista vertailua.

Mukana oli kaksi isoa perhettä, jotka jouduttiin irroittamaan aineistosta. Nämä olivat HYDROLASE- ja OXIDOREDUCTASE-perheet. HYDROLASE-



Kuva 9: Identtisyysvertailun tekevän ohjelmiston arkkitehtuuri. Pääohjelma hallitsee kahta aliohjelmaa. Jokainen ohjelmakomponentti käsittelee tiedostoissa olevaa aineistoa. Suuremmat nelikulmiot ovat ohjelmistokomponentteja ja pienemmät ovat tiedostoja.

perheessä oli yli 400 jäsentä ja se oli laskennallisesti niin raskas, että perhe täytyi pilkkoa pienempiin osiin. Ilman pilkkomista arvioitu laskenta olisi kestänyt yli kymmenen vuorokautta. Perheen pilkkomisesta muodostui 38 osaperhettä. Näistä suurin osa on alle kymmenen jäsenen perheitä. Erikseen on mainittava sellainen HYDROLASE:n osaperhe, jossa ei esiintynyt tarkennetta ollenkaan. Tästä muodostettiin erillinen perhe, jonne jäi n. 150 proteiinia.

Alussa proteiineja oli mukana 2937 kpl. Tästä määrästä karsiutui perheittäisessä identtisyysvertailussa 1090 molekyyliä, joten karsinnan jälkeen aineistoon jäi 1847 proteiiniketjua. Esimerkiksi suuresta HYDROLASE-perheestä, jossa ei ollut tarkennetta, jäi jäljelle noin sata yksilöä. Keskimäärin aineistosta jäi jäljelle noin 60 %, kun identtisyysarvot pääasiassa olivat välillä 30 - 40 %.

Edelleenkin aineistossa saattaa esiintyä päällekkäisyyksiä eli samaan perheeseen kuuluneet joutuivatkin eri perheisiin. Tällainen mahdollisuus saattaa toteutua esimerkiksi silloin, kun perheen nimen ensimmäinen merkkijono on jostain syystä erilainen muihin saman perheen molekyyleihin nähden. Näitä poikkeamia saattaa aiheutua kirjoitusvirheistä tai eri merkintätavoista.

Karsinnasta jäi jäljelle tiedosto, jossa on jokaisen kelpuutetun proteiinimolekyylin identifiointitunnus. Näiden tunnusten avulla haetaan proteiinien sekundaarirakennetiedostot (DSSP). Tiedostot ovat saatavissa Euroopan molekyylibiologian laboratoriosta, joka sijaitsee Saksan Heidelbergissä.

4.3 Rakennetiedostot ja PPII

Tiedostojen siirrossa käytettiin ftp-ohjelmaa, joka hakee .netrc-tiedoston sisältämät tiedostot kohdehakemistosta. Tiedoston koko on rajoitettu 4 kilotavuun, joten id-koodit sisältävää tiedostoa joutui pilkkomaan moneen osaan. Maksimissaan yhteen .netrc-tiedostoon mahtui 233 tiedostoa, joten aineisto jaettiin kahdeksaan osaan. Tiedoston alkuun täytyi liittää sisäänkirjoittautumiskoodit sekä hakemisto-osoite. Haetut tiedostot veivät tilaa 100 Mb.

Kun tiedostoja myöhemmin analysoitiin, muutamat eivät sisältäneet rakenteellista tietoa ollenkaan, vaan näyttivät keskeneräisiltä. Tiedostoja tutkiessa vaikutti siltä, että ajettu ohjelma olisi keskeyttänyt työnsä ensimmäisiin riveihin.

DSSP-tiedostoja sisältävältä palvelimelta ei myöskään löytynyt kaikkia tarvittavia tiedostoja. Halutuista tiedostoista 50 jäi löytymättä. Tiedostojen kohtalo jäi epäselväksi.

4.3.1 DSSP-tiedosto

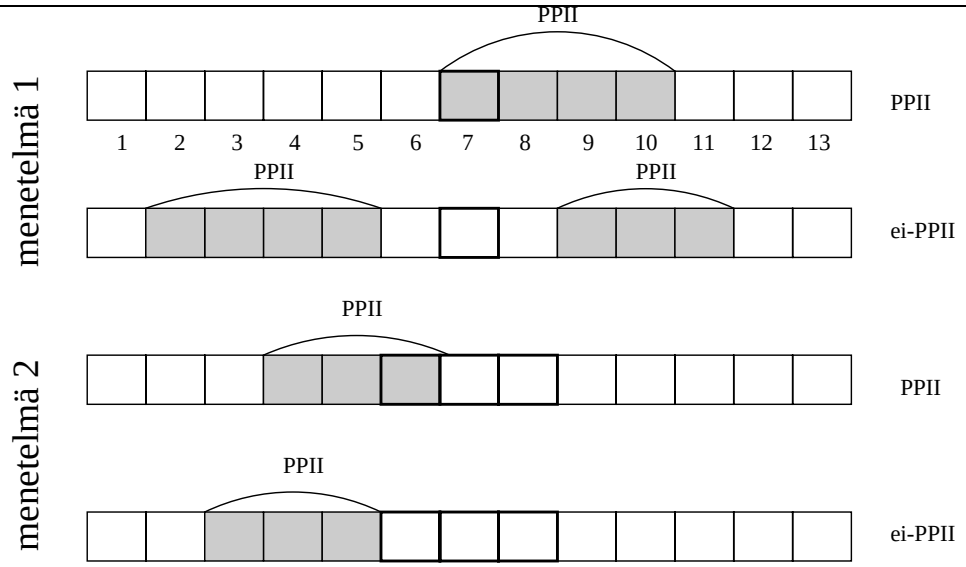
DSSP-tiedostot kuvaavat makromolekyylien rakenteellisia ominaisuuksia. Tiedoston alussa on yleistä tietoa ohjelman kehittäjistä ja molekyylistä. Varsinaisen proteiiniketjun alkaminen ilmoitetaan #-merkillä. Tämän merkin jälkeen seuraavilla riveillä ovat ketjun aminohapot; yhdellä rivillä aina yhden tiedot. Tiedosto saattaa olla proteiinin pituuden mukaan muutaman rivin mittaisesta muutamiin tuhansiin riveihin.

Yhdellä rivillä ilmoitetaan yksittäisen aminohapon järjestysnumero ketjussa, kirjaintunnus, tunnettu sekundaarirakenneluokka, keskushiiliatomin sijainti kolmella koordinaatilla, taipumiskulmat ϕ ja ψ sekä paljon muuta informaatiota.

Tiedostoissa tunnetaan muutama yleisin sekundaarirakenne: α-kierre (H), β-tikapuut ($B + E$), 3_{10}-kierre (G) sekä β-käänne (B). Näistä erityisesti oikeankätistä α-kierrettä esiintyy proteiineissa runsaasti [1].

4.3.2 Opetus- ja testiaineistojen muodostaminen

Ohjelma, jolla DSSP-tiedostojen sekundaarirakenteet on määrätty, perustuu hahmontunnistukseen [12]. Tämä viittaa siihen, että sekundaarirakenteiden



Kuva 10: Ikkunointimenetelmät 1 ja 2 13-mittaisella tarkasteluikkunalla. Harmaalla merkityt laatikot ovat paikannettuja PPII-rakenteita ja valkoiset laatikot ovat aminohappoja, joiden kohdalla ei PPII-rakennetta esiinny. Ikkunointimenetelmä 1 hyväksyy ikkunan PPII-luokkaan, jos rakenne on keskimmäisen aminohapon kohdalla. Ikkunointimenetelmä 2 hyväksyy sekvenssin PPII-luokkaan, jos rakenne esiintyy kolmen keskimmäisen aminohapon kohdalla. Muutoin ikkunan sisältämä sekvenssi kuuluu luokkaan ei-PPII.

rajat eivät ole tarkkoja ja rakenteiden säännöllisyys ei myöskään ole kovin tarkka. Polyproliini II-rakenteen etsintään käytetään kuitenkin tarkkoja algoritmisia ohjeita Adzhubein ja Sternbergin artikkelista (katso lähde [1]).

Opetusryhmä-nimikettä käytetään karsinnoista jääneeseen tiedostojoukkoon, josta on erotettu 10 % testiaineistoon. Loput 90 % kuuluu opetusaineistoon. Testijoukko erotetaan systemaattisella otannalla, jossa joka kymmenes poimitaan testiaineistoon. Tällöin koko aineistosta voidaan irroittaa kymmenen erilaista opetusryhmää. Yhdelle opetusryhmälle tehdään neljä etsintäajoa, jolloin etsitään kahden eri ikkunointiratkaisun tapaukset ja kahden eri mittaisen tarkasteluikkunan tapaukset.

Tarkasteluikkuna on tietyn kokoinen “sapluuna”, jota liu'utetaan proteiinisekvenssin ylitse. Ikkunoinnilla tarkoitetaan menetelmää, jossa sekvenssit valitaan neuroverkolle. Ikkunointi määrää, kuuluuko tarkastelun kohteena oleva sekvenssin osa PPII- tai ei-PPII-luokkaan. Tällöin tarkastellaan PPII-rakenteen sijaintia suhteessa tarkasteluikkunan keskiosaan (katso kuva 10).

Ikkunan pituutta harkittaessa täytyy pohtia, mitä sekvenssin pituus vaikut-

taa ennustustarkkuuteen. Muutamissa artikkeleissa pohditaan tarkasteluikkunan pituutta, kun ennustuksen kohteena on α- ja β -rakenteet. Ruggiero, Sacile ja Rauch ovat tutkimuksessaan saaneet parhaat tulokset 13-mittaisella tarkasteluikkunalla [24]. Raportissa myös mainitaan toisen tutkijaryhmän saaneen parhaat tulokset 17-mittaisella tarkasteluikkunalla.

Ikkunan pituutta kasvatettaessa yhdellä lisääntyvät neuroverkon yhteydet määrällä 20 $\times$ ensimmäisen piilokerroksen solmujen lukumäärä. Tämä taas vaikuttaa siihen, että opetusaineiston kokovaatimus kasvaa varsin kovaa vauhtia. Tämän työn yhteydessä päädyttiin käyttämään tarkasteluikkunan kokoina pituuksia 7 ja 13.

Aluksi työssä vertaillaan ikkunointimenetelmiä 1 ja 2 (kuva 10) sekä selvitetään, mitä vaikutusta PPII:n tapauksessa on tarkasteluikkunan pituudella. Koska ikkunointimenetelmällä 1 saadaan yli 8000 opetusalkiota ja ikkunointimenetelmällä 2 yli 14000 opetusalkiota (ei riipu olennaisesti tarkasteluikkunan pituudesta), määräävät nämä luvut suurimman mahdollisen tarkasteluikkunan koon sekä verkon piilosolmujen maksimimäärän.

Opetusaineistoa täytyisi olla (nyrkkisääntöjen mukaan) noin kymmenkertainen määrä neuroverkon yhteyksiin nähden. Oletetaan, että tiedetään seuraavat asiat: käytössä on kaksi aineistoa, jossa toisessa on alle 15000 ja toisessa alle 9000 opetustapausta; yksi tarkasteluikkunan positio laajenee 20:een mahdolliseen aminohappoon; tarkasteluikkunan pituus on l; piilosolmujen lukumäärä on p ja tulossolmujen lukumäärä on 2. Tällöin muuttujien välille saadaan yhteydet

$$10(20lp + 2p) = 9000 \tag{5}$$

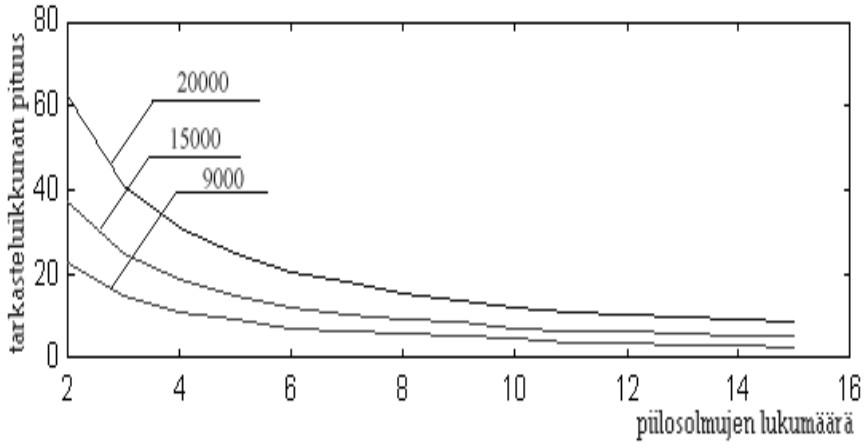
ja

$$10(20lp + 2p) = 15000. \tag{6}$$

Yhteyksiä voidaan havainnollistaa graafisesti (katso kuva 11). Kuvasta nähdään piilosolmujen määrät, kun tiedetään ikkunan pituus ja opetusaineiston määrät. Sallitut piilosolmujen määrät ja tarkasteluikkunoiden koot pitää valita aineiston koon mukaan lasketun käyrän alapuolelta.

4.3.3 Polyproliini II-rakenteen paikantaminen ja proteiinin ikkunointi

PPII-rakenteiden etsinnässä käytetään tutkijoiden Adzhubei ja Sternberg kehittämiä menetelmiä sekä vertaillaan tuloksia heidän saamiinsa tuloksiin [1].



Kuva 11: Piilosolmujen ja tarkasteluikkunan väliset enimmäisrajat tietyn kokoisille opetusjoukoille.

He ovat artikkelissaan analysoineet 80 proteiinia, joista löytyi 96 säännöllistä PPII-rakennetta.

Rakenteiden etsinnän ensimmäinen ehto asetetaan virtuaaliselle α-kulmalle. Tämä kulma toimii seulana, joka karsii ne rakenteet joukosta, joiden ϕ- ja ψ-kulmat eivät kuulu sallitulle alueelle tai kätisyys on väärä. Kulma α lasketaan kaavalla

$$\alpha = 180^\circ + \psi_i + \phi_{i+1} + 20^\circ(sin\ \phi_i + sin\ \psi_{i+1}).$$

Polyproliini II-rakenteen suurin esiintymä keskittyy pisteen $\alpha = -110^\circ$, $\phi = -75^\circ$ ja $\psi = 145^\circ$ ympäristöön. Geometrisesti nämä kulmat vastaavat rakennetta, jossa on kolme aminohappoa kierroksella. Tästä muodostuu kolmiomainen avaruudessa etenevä rakenne. Kuvassa 12 nähdään osa $\phi\psi$-avaruudesta. Kuvasta on mahdollista tarkastella PPII-rakenteelle asetettua ehtoa: $-145^\circ < \alpha < -70^\circ$ (harmaa alue). Kulmista ϕ- ja ψ tarkastellaan rakenteen säännöllisyyttä. Rakenteiden etsimisessä meneteltiin algoritmin 3 mukaisesti.

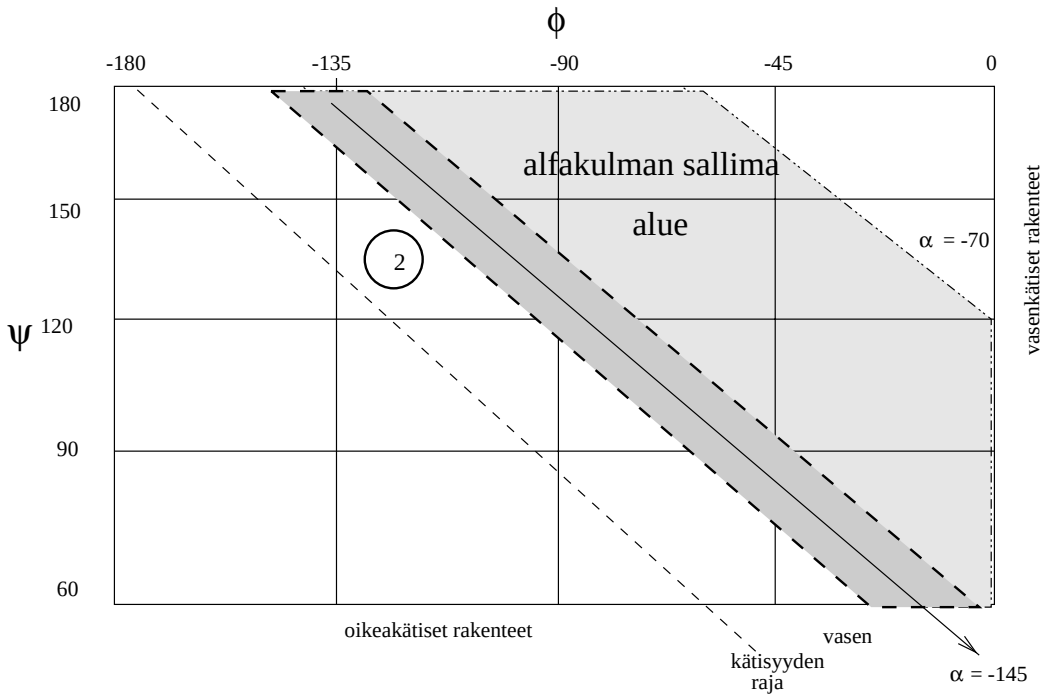
Rakenteen säännöllisyys lasketaan kaavalla

$$D = (d_{1,2} + d_{2,3} + d_{3,4} + ... + d_{n-1,n})/n, \tag{7}$$

missä

$$d_{k-1,k} = \sqrt{(\psi_{i-1} - \psi_i)^2 + (\phi_i - \phi_{i+1})^2}. \tag{8}$$

Säännöllisyys on peräkkäisten kulmien ϕ ja ψ avulla muodostetun etäisyyden keskiarvo. Algoritmissa lasketaan havaitun PPII-kandidaatin säännöllisyys



Kuva 12: $\phi\psi$-avaruus ja PPII-rakenteen sijainti. PPII-rakenteen etsinnässä rajataan α-kulma harmaaseen alueeseen. Tämän jälkeen tehdään koko rakenteen ja sen osavälien säännöllisyystarkastus. Pisteessä 2 sijaitsee β-tikapuut -rakenteen tihein esiintymä.

koko matkalta. Tämän jälkeen jokaiselle mahdolliselle osavälille lasketaan vastaava säännöllisyysarvo.

Kuvassa 12 näkyy toinenkin säännöllinen sekundaarirakenne β-tikapuut (β-ladders), joka on merkitty numerolla 2. PPII- ja β-tikapuut -rakenteet sijaitsevat lähellä toisiaan. Näiden kahden rakenteen toinen raja on kaistaleessa $\alpha = -145 \pm 10°$. Tämä erottuu kuvassa 12 tummana vinona kaistaleena, jonka keskellä menee raja $\alpha = -145$.

Rakenteita kuvaavia kulmia on esitelty kuvassa 13. Algoritmissa käytetty parametri *pituus* vaikuttaa siihen, montako sellaista aminohappoa täytyy peräkkäin esiintyä, joiden kulmat ψ_{i-1}, ψ_i ja ϕ_i, ϕ_{i+1} täyttävät säännöllisyysehdot. Ehtoja kokeiltiin kahta: pituudet 2 ja 3. Adzhubei ja Sternberg ovat nähtävästi käyttäneet pituutta 2, sillä tällä arvolla löytyi enemmän rakenteita. Koska pituudella 2 löytyy paljon rakenteita, muuttuu opetusjoukon hallinta raskaaksi. Tämän takia lyhyemmän rakenteen aineistoa testataan vain yhdellä opetusryhmällä.

Algoritmi 4 tallettaa jokaisen PPII-luokkaan kuuluvan ikkunallisen, mutta tallettaa järjestyksessä vain joka k:nnen ei-PPII-luokan ikkunallisen. Muu-

Algoritmi 3 Rakenteiden etsintä, parametrina pituus

```
1: while proteiinimolekyylejä jäljellä do
2:   while aminohappoja jäljellä do
3:     if alaraja < α < yläraja then
4:       while alaraja < α < yläraja do
5:         ota seuraava aminohappo käsittelyyn
6:       end while
7:       if peräkkäisiä rakenteita enemmän kuin pituus then
8:         ketjun säännöllisyystarkastus (D < 50)
9:         if rakenne säännöllinen ja rakenteessa enemmän aminohappoja
           kuin pituus then
10:          tarkastetaan jokaisen osavälin säännöllisyys
11:          if jokaisen osavälin säännöllisyys D < 50 then
12:            merkitään välin aminohapot PPII-aktiivisiksi
13:          end if
14:        end if
15:      end if
16:    end if
17:    ota seuraava aminohappo käsittelyyn
18:  end while
19:  proteiinin ikkunointi (algoritmi 4)
20:  ota seuraava proteiini käsittelyyn
21: end while
```

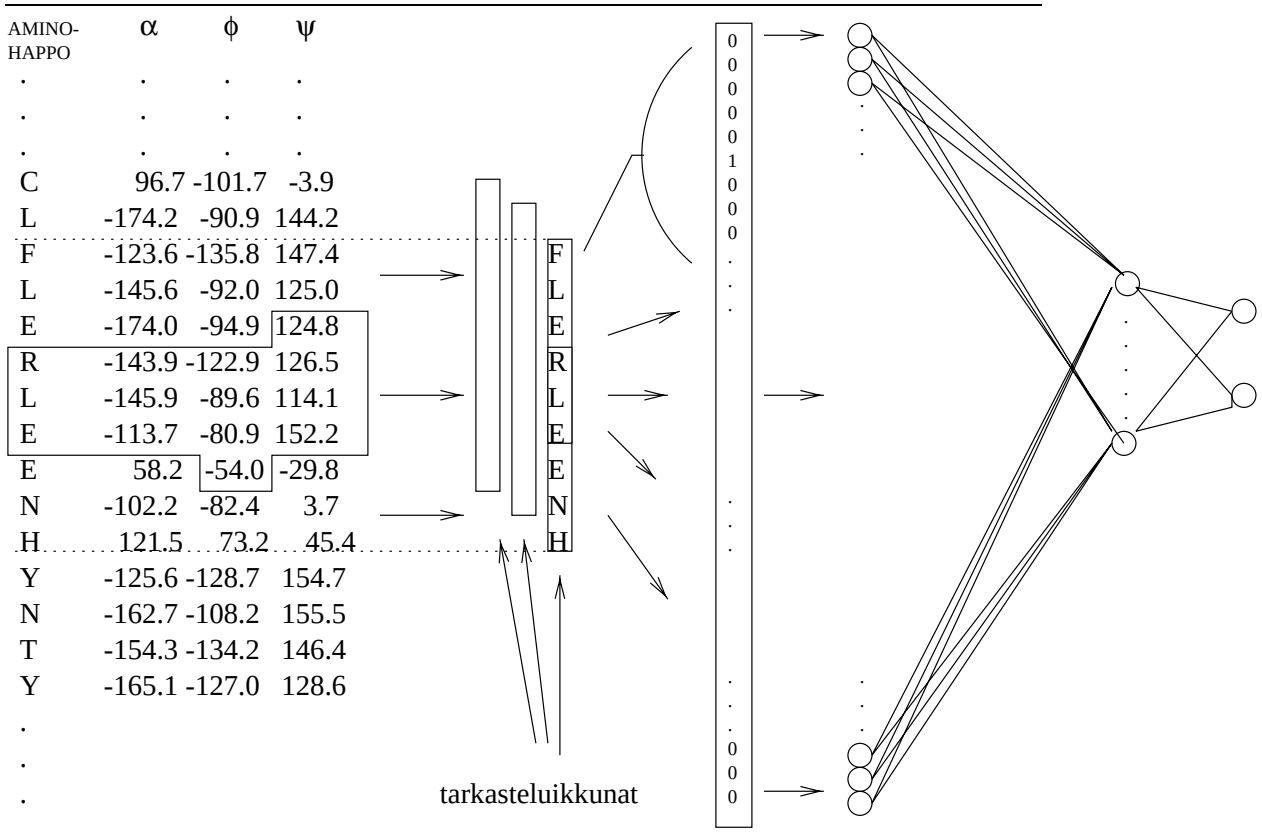
toin ei-PPII-luokkaan tulisi moninkertainen määrä alkioita.

Etsintäoperaatiossa proteiini selataan läpi kahdesti alusta loppuun. Ensimmäisellä kerralla paikannetaan PPII-rakenteet ja toisella kerralla ikkunoidaan proteiini. Ikkunoinnissa tarkastellaan kuvan 10 mukaisia ehtoja. Jos yhden ikkunan ehdot täyttyvät, tallennetaan ikkuna PPII-aineistoon, muutoin ei-PPII-aineistoon.

Kun kaikki proteiinit on käsitelty, on syntynyt kaksi tiedostoa. Toisessa on sekvenssejä PPII-luokasta ja toisessa sekvenssejä ei-PPII-luokasta. Sekvenssit ovat tiedostossa riveittäin siten, että yhden tarkasteluikkunan sisältämät aminohapot ovat yhdellä rivillä.

4.4 Aineiston ominaisuuksia

Rakenteen pituuden määrittely jää Adzhubein ja Sternbergin artikkelissa hieman epäselväksi. Ongelmaksi jää, mistä rakenne alkaa ja mihin se loppuu.



Kuva 13: Sekvenssin muuttuminen neuroverkon ymmärtämään muotoon. Rakennetiedoston kulmat (monikulmiossa) osoittavat PPII-rakennetta. Aminohapot R, L, E asetetaan PPII-aktiivisiksi. Tämän jälkeen tarkasteluikkunan avulla kerätään sekvenssit PPII- ja ei-PPII-luokkiin. Yhdestä tarkasteluikkunasta saadaan pitkä nollia ja ykkösiä sisältävä vektori, joka voidaan antaa syötteenä neuroverkolle.

Tämän työn yhteydessä rakenteen alku määrätään sijaitsemaan sen aminohapon i kohdalla, jolla säännöllisyysehdot täyttävä ψ-kulma kohdassa $i-1$. Rakenteen loppu on sen aminohapon k kohdalla, jolla on säännöllisyysehdon täyttävä ϕ-kulma kohdassa $k+1$ (katso kuva 13).

Kun rakenteen alku ja loppu määritellään näin, tulee epäselvyyttä, tarkoitetaanko artikkelissa mainitulla pituudella samaa asiaa. Tämän takia aineisto on tuotettu kahdella eri tavalla. Kun rakenteelle vaaditaan kolme aminohappoa, jossa kulmat ovat määriteltyjen ehtojen rajoissa, saadaan artikkelin määrästä vain puolet rakenteita. Kun aminohappoja vaaditaan vain kaksi, saadaan puolestaan hieman enemmän kuin heidän artikkelissaan (katso taulukko 1).

Suurimmassa osassa aineiston ominaisuuksien mittauksia käytetään ik-

Algoritmi 4 proteiinin ikkunointi

```
1: asetetaan tarkasteluikkuna proteiinin alkuun siten, että ikkunan ensimmäinen ruutu on ensimmäisen aminohapon kohdalla ja viimeinen tarkasteluikkunan pituuden verran eteenpäin
2: while tarkasteluikkunan viimeinen alkio ei ole proteiinin lopussa do
3:    if PPII-rakennetta esiintyy ikkunointiehdon mukaisesti ikkunan alueella then
4:       talletetaan PPII-tiedostoon ikkunan kohdalla oleva sekvenssi
5:    else
6:       if ei-PPII-rakenteen järjestysnumero on jaollinen luvulla k then
7:          talletetaan ei-PPII-tiedostoon ikkunan kohdalla oleva sekvenssi
8:       end if
9:    end if
10:   siirretään tarkasteluikkunaa yhden aminohapon verran eteenpäin
11: end while
```

kunointimenetelmää 1 ja sekvenssin pituutta 13. Tämä tehdään näin, koska ikkunointimenetelmää 1 on perinteisesti käytetty sekundaariennustuksissa ja pituus 13 on yläraja aineiston määrän takia.

4.4.1 PPII-rakenteiden esiintymisestä

Rakenteen pituudella 3 PPII-esiintymiä löydetään keskimäärin 1.26 %. Hajonta on aika suuri, sillä mukana on proteiineja, joissa ei esiinny PPII-rakennetta ja joissakin osuus on yli 10 %. Analyysi löysi yhdeksän proteiinia, joissa PPII-esiintymiä on yli 10 %. Proteiineja löytyi PPII-rakenteen esiintymistiheydellä 5 - 10 % 71 kpl., tiheydellä 2 - 5 % 319 kpl. ja tiheydellä 0 - 2 % löytyi 463 kpl. Kokonaisuudessaan PPII-rakennetta esiintyy 862:ssa proteiinissa, kun kaikkiaan aineistossa on mukana 1849 proteiinia. Proteiinit, joissa esiintyi runsaimmin PPII-rakennetta, on esitetty liitteessä D.

Rakenteen pituudella 2 PPII:n esiintymistiheys on n. 3 %. Tästä voidaan päätellä, että noin puolet PPII-rakenteista on kahden aminohapon mittaisia.

4.4.2 Lukumääriä ja suhteita

Käsitellään aluksi rakenteen pituudella 3 saatuja lukumääriä. Ikkunointimenetelmä 1 löytää opetusaineistosta keskimäärin 6950 PPII-rakennetta ja ikkunointimenetelmä 2 löytää keskimäärin 11000 PPII-rakennetta.

Ikkunointiratkaisulla 1 ja tarkasteluikkunan pituudella 7 saadaan keskimäärin 7000 PPII-tapausta, kun 13-mittainen tarkasteluikkuna tuottaa

Taulukko 1. Työssä käytettyjen algoritmien löytämät PPII-rakenteet suhteessa Adzhubein ja Sternbergin tuloksiin [1]. Tähdellä merkityt ovat multimeerejä.

Proteiini	artikkeli	pituus 2	pituus 3
1CC5	1	1	0
1HIP	2	2	0
1PPT	1	0	0
2ACT	2	6	1
2APR	3	6	1
2CCY*	1	2	2
2CDV	2	2	1
2CYP	5	6	2
2SNS	1	1	0
3BCL	4	6	3
3BLM	1	9	2
3GRS	4	7	0
3EST	4	8	4
4CPV	1	1	1
4DFR*	3	5	0
4MDH*	2	5	0
8ABP	1	2	1
9WGA*	4	9	1

keskimäärin 6800 PPII-tapausta. Vastaava suhde esiintyy myös ikkunointiratkaisulla 2.

Kun rakenteet on etsitty ja kirjoitettu tiedostoon, poistetaan identtiset tapaukset kummankin luokan sisältä. Identtisyyksien takia PPII-luokasta voidaan poistaa keskimäärin 37 % tapauksista ja ei-PPII-luokan tapauksista vain hieman yli 10 %. Tulos kertoo siitä, että PPII-luokan tapauksissa on enemmän samankaltaisuutta kuin ei-PPII-luokassa.

Rakenteen pituudella 2 ja ikkunointimenetelmällä 1 löydetään kaikkiaan n. 18000 PPII-tapausta. Kun näistä poistetaan identtiset, jää jäljelle n. 11000 tapausta. Aineistosta kuitenkin poistetaan testijoukot ennen opetusta, jolloin opetukseen jää n. 10000 PPII-tapausta. Näin isolla opetusjoukolla ei tietokoneen keskusmuistimäärä riitä ja opetus muuttuu hitaaksi.

4.4.3 Frekvenssit

Frekvenssiä koskevat tulokset on saatu tarkasteluikkunan pituudella 13 ja ikkunointimenetelmällä 1. Taulukossa 2 nähdään analyysin tulokset. Liitteessä E on esitetty PPII-sekvenssien ja ei-PPII-sekvenssien aminohappojen suhteet tarkasteluikkunan pituudella 13.

Taulukko 2. PPII- ja ei-PPII-luokissa esiintyneet aminohappojen frekvenssit. Analyysi on tehty 13-mittaisella tarkasteluikkunalla, joten rivejä on 13 kpl. Sarakkeet kuvaavat DSSP-tiedostossa olleita kirjaimia (sarakkeita on siis enemmän kuin proteiineissa esiintyviä aminohappoja).

PPII

A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	X	Y	Z
754	27	95	624	568	372	969	235	552	584	758	178	545	596	353	510	729	588	689	138	8	394	1
729	25	89	607	551	359	1025	222	539	624	724	156	547	645	396	489	726	593	631	156	9	440	1
690	21	101	580	525	399	1110	211	550	636	707	170	532	586	390	504	662	615	688	145	9	418	1
644	23	94	511	515	465	1106	195	617	634	737	174	467	516	369	488	623	664	823	151	13	419	1
714	39	80	382	567	504	828	231	634	668	769	186	365	594	363	531	582	715	969	144	12	408	1
742	41	87	291	559	502	313	205	718	622	1006	198	294	1081	367	519	466	687	1074	131	9	424	0
723	30	83	538	487	358	277	175	580	544	948	153	364	1733	299	438	673	668	841	113	6	305	2
854	22	63	691	544	245	333	159	444	578	847	132	427	1850	336	441	854	573	639	97	8	198	2
754	17	69	823	698	305	706	178	394	585	698	128	553	1117	378	451	822	583	614	117	8	283	0
715	22	90	916	802	338	984	213	333	551	562	142	584	649	436	398	798	707	541	107	12	346	0
727	22	101	739	740	374	920	214	444	534	663	183	539	591	454	411	785	697	609	129	11	364	0
751	28	102	641	681	354	798	244	527	604	753	211	521	608	411	446	708	662	689	153	10	366	0
827	28	102	687	661	391	714	228	520	602	772	196	477	573	402	451	674	619	716	206	10	425	1

ei-PPII

A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	X	Y	Z
870	19	114	625	639	425	827	202	555	616	898	200	490	455	394	429	599	608	789	139	2	427	0
854	16	98	611	624	448	874	237	623	593	806	219	476	447	400	456	619	626	734	150	3	378	2
859	24	96	589	665	421	860	214	564	599	818	211	533	463	344	447	644	638	783	152	6	383	0
915	16	116	594	633	400	815	213	560	627	901	214	504	460	345	445	658	608	752	127	7	400	0
866	10	95	621	676	394	878	232	543	634	864	235	495	437	365	462	678	620	651	128	4	419	0
913	31	97	623	620	399	897	227	598	653	783	225	525	424	372	455	637	632	697	142	5	362	0
860	17	101	652	631	395	891	233	559	613	818	184	461	444	397	493	695	655	689	136	9	369	1
854	14	97	587	672	399	850	243	591	638	898	183	461	440	365	501	650	566	719	186	6	390	0
868	19	88	603	613	412	871	218	614	597	854	246	481	425	349	447	660	630	762	140	7	408	0
853	12	119	603	648	391	831	225	594	602	880	225	469	444	372	465	638	595	762	151	7	419	1
833	18	89	636	635	415	830	201	563	648	885	221	503	433	403	463	588	622	758	159	6	398	0
848	12	112	620	669	424	824	259	562	636	800	203	482	468	395	476	641	624	706	170	3	361	2
885	15	91	640	624	409	821	201	589	622	824	228	517	444	369	464	635	628	787	144	6	366	0

Huomioitavaa on PPII-joukon keskimmäiset rivit, joissa sijaitsevat merkittävimmät eroavaisuudet. Rakenteelle tärkeitä aminohappoja on tummennettu niistä kohdista, missä frekvenssiarvot ovat poikkeavat. Kiinnostavia aminohappoja ovat ainakin G, H, L, N, P, S, V ja Y.

4.4.4 Luokkien välinen Hamming-etäisyys

Hamming-etäisyys lasketaan aineiston pituuksille 5, 7 sekä 13. Aineistot on saatu ikkunointimenetelmällä 1. Joukkojen väliset samankaltaisuudet voidaan paljastaa etsimällä jokaiselle PPII-luokan tapaukselle lähin sukulainen vastakkaisesta luokasta. Lähimmistä etäisyyksistä voidaan nähdä, kuinka paljon luokat ovat päällekkäin.

Hamming-etäisyys on metriikka, joka on määritelty N-ulotteisille bittivektoreille. Kahden vektorin etäisyys on niiden positioiden lukumäärä, missä vektoreilla on eri arvot [10]. Etäisyys voidaan laskea myös kaavalla

$$h = \sum_{i=1}^{n} (x_i(1 - y_i) + y_i(1 - x_i)), \tag{9}$$

missä x_1 on toisen vektorin muuttuja i ja y_i on vastakkaisesta luokasta olevan vektorin muuttuja i.

Kun jokaiselle PPII-luokan yksilölle lasketaan Hamming-etäisyys jokaisen ei-PPII-yksilön kanssa ja valitaan näistä etäisyyksistä pienin, saadaan taulukon 3 mukaiset tulokset.

Taulukko 3. Hamming-etäisyydet tarkasteluikkunan pituuksilla 5, 7 ja 13. Arvot on esitetty prosentteina. Katkoviivan yläpuolella on PPII-sekvenssien etäisyydet ei-PPII-luokan sekvensseistä. Katkoviivan alapuolelle on PPII-luokan sekvenssien etäisyydet oman luokan sekvensseistä. Pituussarakkeessa oleva merkintä n2 ja n3 tarkoittaa rakenteen pituutta.

tarkasteluikkunan pituus	etäisyys 0	1	2	3	4	5	6	7	8	9
5 n3	1	19	79	1	0	0	0	0	0	0
7 n3	1	1	4	62	32	0	0	0	0	0
13 n3	1	1	0	0	0	0	1	15	67	15
13 n2	1	0.5	0	0	0	0.5	3	33	60	2
13 n3	7	4	4	3	4	0	6	22	45	5
13 n2	8	4	4	3	4	0	8	32	35	1

Taulukosta nähdään, että tarkasteluikkunan pituudella on suuri vaikutus moodin sijaintiin. Viiden aminohapon mittaisella sekvenssillä on eri luokkien lähimmät tapaukset yleisimmin kahden aminohapon päässä. Tarkasteluikkunan pituudella 13 lähimpien alkioiden moodi sijaitsee jo kahdeksan aminohapon päässä.

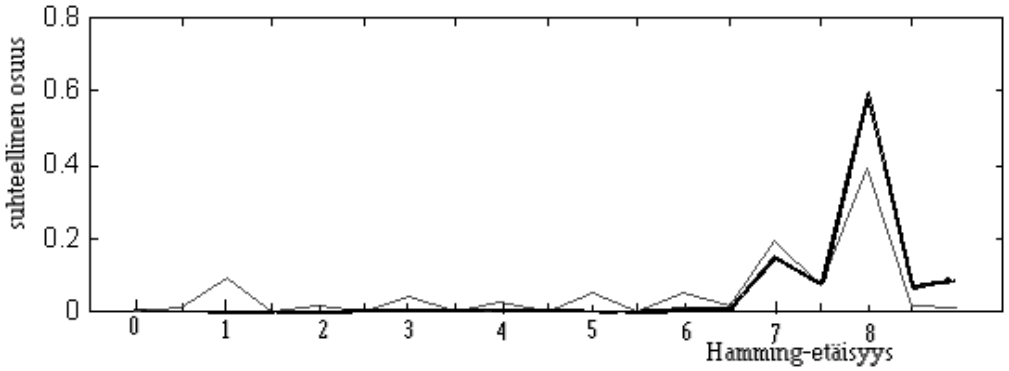
Kun PPII-luokan tapauksia verrataan oman luokan tapauksiin, on moodi yhtä etäällä kuin vertailussa ei-PPII-luokan kanssa. Oman luokan sisällä on kuitenkin enemmän sellaisia tapauksia, jotka ovat alle 7 etäisyydellä. Ilmiö on selvästi huomattavissa myös kuvasta 14.

Etäisyyttä on tarkasteltu hieman toisin kuvassa 15. Aminohappojen välillä on biokemiallisia samankaltaisuuksia. Sukulaisuudet saadaan taulukosta PAM250 [17] (aiheesta on luvussa 4.2). Kun sukulaisuudet otetaan huomioon, saadaan opetusaineistolle uusi *sukulaisuusmetriikka*.

Kahden sekvenssin välinen sukulaisuus muodostetaan siten, että verrataan sekvenssien samoissa positioissa olevia aminohappoja toisiinsa. Aminohappoparin sukulaisuusarvoksi otetaan PAM250-taulukon antama sukulaisuusarvo. Kahden sekvenssin sukulaisuusarvoksi summataan kaikkien aminohappoparien antamat arvot yhteen.

Sukulaisuusmetriikan takia PPII-luokan sisällä saaduissa tuloksissa "häntä" on keskittymän oikealla puolella. Kuvista nähdään, että PPII-luokan moodi on hieman oikealla ei-PPII-luokan vastaavaan verrattuna. Menetelmä ei kuitenkaan muuta oleellisesti PPII- ja ei-PPII-luokan välistä suhdetta.

Viimeiseksi tarkastellaan satunnaisesti valitun opetusjoukon ja tämän testaukseen erotetun testijoukon välistä Hamming-etäisyyttä. Tulokset on esi-



Kuva 14: Aineiston Hamming-etäisyyksien histogrammit sekvenssin pituudelle 13. Paksumpi kuvaaja esittää PPII- ja ei-PPII-luokkien välillä laskettuja Hamming-etäisyyksiä. Ohuempi kuvaaja esittää PPII-luokan sisällä laskettuja etäisyyksiä.

tetty kuvassa 16.

Histogrammista ja frekvenssiluvuista paljastuu voimakkaana identtisten tapausten joukko (etäisyys 0), jonka osuus on koko aineistosta 15.7 %. Toinen iso rykelmä keskittyy etäisyydelle kahdeksan. Etäisyys seitsemän ja kahdeksan välissä on seurausta DSSP-tiedoston ylimääräisistä merkeistä.

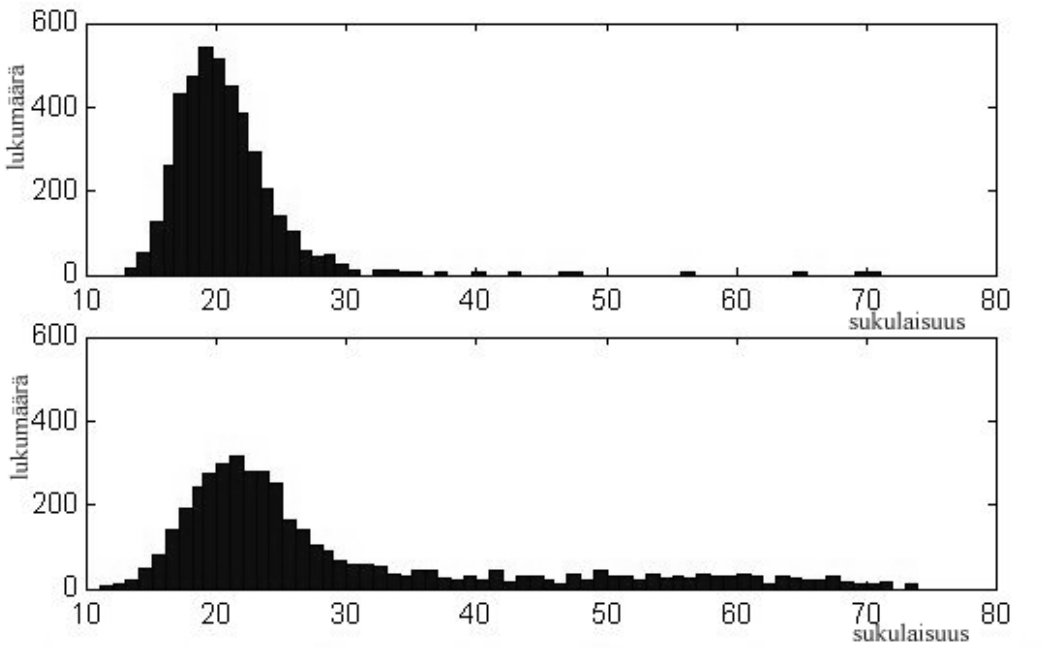
4.4.5 Aineiston opittavuus

Swingler on kirjassaan esitellyt menetelmiä, joiden avulla voidaan tutkia aineiston opittavuutta [26]. Menetelmät perustuvat Shannonin informaatioteoriaan. Aineistoa tarkastellaan entropian, ehdollisen entropian ja keskinäisinformaation odotusarvon avulla.

Laskennassa käytetään yhtä satunnaisesti valittua opetusryhmää, josta testiaineisto on poistettu. Aineistossa on 8500 tapausta. Luokkien sisällä ei ole identtisiä alkioita, mutta vastakkaisissa luokissa esiintyy yhteensä 20 identtistä paria. Tällöin vain toisessa luokassa olevia sekvenssejä on 8460 ja molemmissa luokissa olevia sekvenssejä on 20 paria. Merkitään, että X tarkoittaa syötetapauksia ja Y tulostapauksia. Aluksi syöte- ja tulostapauksille lasketaan entropia-arvot kaavalla

$$H = \sum_{i=1}^{n} P_i \log \frac{1}{P_i}, \tag{10}$$

missä P_i on tapauksen i todennäköisyys ja logaritmina käytetään luonnollista logaritmia. Syötevektoreille vain toisessa luokassa esiintyvien tapauksien todennäköisyys on 1/8500 ja kummassakin luokassa esiintyvien todennäköisyys



Kuva 15: Aineiston sukulaisuuksien histogrammit. Laskennasssa on käytetty apuna aminohappojen sukulaisuuksia. Vaaka-akselilla on laskettu sukulaisuusarvo ja pystyakselilla on esiintymien lukumäärät. Ylemmässä kuvassa PPII-tapauksia on verrattu ei-PPII-tapauksiin. Alemmassa kuvaajassa on sukulaisuudet laskettu PPII-luokan sisällä.

on 2/8500. Syötetapausten entoropia-arvo on tällöin

$$H(X) = \sum_{i=1}^{8460} \frac{1}{8500} \log \frac{1}{\frac{1}{8500}} + \sum_{i=1}^{20} \frac{2}{8500} \log \frac{1}{\frac{2}{8500}} \approx 9.01 \approx \log 8500. \qquad (11)$$

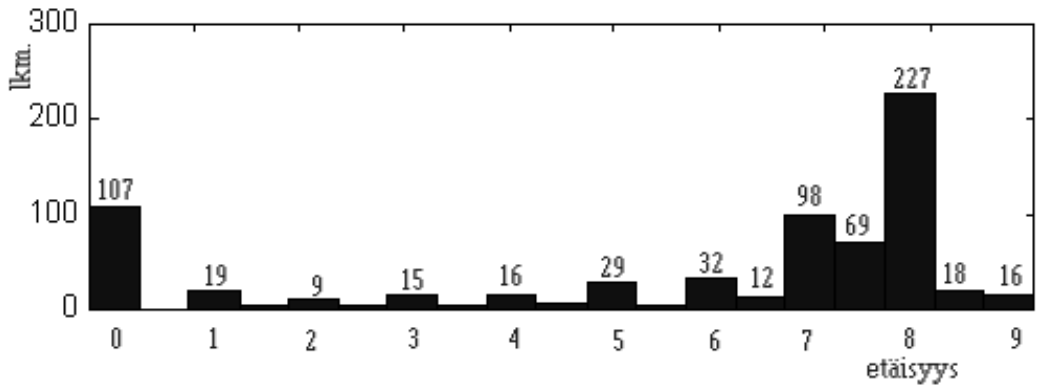
Tulosluokkia on kaksi ja kummankin luokan todennäköisyys on 1/2. Tällöin tulosluokille saadaan entropia-arvo

$$H(Y) = \sum_{i=1}^{2} \frac{1}{2} \log \frac{1}{\frac{1}{2}} = \log 2. \qquad (12)$$

Arvo $H(Y)$ on tulosluokan entropian maksimiarvo ja $H(X)$ on melkein entropian maksimiarvo $\log 8500$, joten opetusjoukon voidaan sanoa olevan hyvin tasapainossa [26].

Ehdollinen entropia tarkastelee luokan y_i entropiaa, kun se kuuluu syötteelle x_j. Tämä määritellään kaavalla

$$H(Y|X) = \sum_{i=1}^{n} \sum_{j=1}^{m} P(y_i, x_j) \log \frac{P(x_j)}{P(y_i, x_j)}, \qquad (13)$$



Kuva 16: Satunnaisesti valitun opetusjoukon ja tämän testijoukon väliset Hamming-etäisyydet PPII-luokan sisällä.

missä $P(y_i, x_j)$ kuvaa todennäköisyyttä, että luokka y_i ja tapaus x_j tapahtuvat molemmat.

Niissä tapauksissa, joissa alkio esiintyy vain toisessa joukossa (8460 kpl.), on $P(y_i, x_j) = P(x_j)$, joten logaritmi antaa arvon 0. Niitä tapauksia, joissa sama alkio esiintyy kummassakin luokassa on 20 kpl. ja yhtä tällaista tapausta on aina 2 kpl., joten

$$P(y_i, x_j) = \frac{1}{2}\frac{2}{8500} \tag{14}$$

ja

$$P(x_j) = \frac{2}{8500}. \tag{15}$$

Tapauksia on yhteensä 40, joten ehdolliselle entropialle saadaan kokonaisarvo

$$40 * (\frac{1}{8500} \log \frac{\frac{2}{8500}}{\frac{1}{8500}}) \approx 0.0033. \tag{16}$$

Ehdollinen entropia antaa opetusjoukolle todella pienen tuloksen. Tällöin $H(Y|X) : H(Y) \approx 0.0048$. Tulos on välillä $(0, 1)$ lähellä lukua 0. Tämä on Swinglerin mukaan hyvän opittavuuden merkki.

Molemminpuolinen informaatio määritellään kaavalla $I(X;Y) = H(X) - H(X|Y)$. Informaation ominaisuuksien mukaan $I(X;Y) = I(Y;X)$ [26], jolloin $I(Y;X) = H(Y) - H(Y|X)$. Nyt voidaan käyttää aikaisempia tuloksia hyväksi, jolloin saadaan, että $I(Y;X) = \log 2 - 0.003 \approx 0.69$.

Swingler on esittänyt, että molemminpuolisen informaation ja tulostapausten entropian suhde paljastaa aineiston opittavuuden. Suhteelle saadaan

arvo 0.995. Arvot voivat olla väliltä $(0, 1)$, joten aineisto antaa voimakkaita viitteitä hyvään opittavuuteen.

Entropia-tarkastelun aineistolle antamat tulokset viittaavat todella hyvään oppimiseen. Tämä johtunee lähinnä siitä, että vastakkaisten luokkien identtisiä on aineistossa vähän.

4.4.6 Aineiston ongelmista

PPII:n vähäinen esiintyminen proteiineissa on vaikea ongelma. Miten pitäisi menetellä opetus- ja testijoukon muodostamisessa ja ennustustarkkuutta määriteltäessä?

Kirjallisuudessa on pohdittu eri näkökulmista opetusjoukon ongelmia. Erikokoisten opetusjoukkojen tapauksesta käytetään termiä epätasapainossa olevat luokat. Yleisimpiä ratkaisuja tällöin ovat dominoivan luokan karsinta [2], pienemmän luokan monistus [2], [14], yliopetus [14] sekä keinotekoisen datajoukon muodostaminen [14].

PPII-aineistoa saatiin niukasti rakenteen pituudella 3 ja sopivasti pituudella 2. Kummassakin tapauksessa PPII-joukosta muodostui mitättömän pieni verrattuna ei-PPII-luokkaan.

Jos pienemmän luokan alkioita monistetaan, tulee aineistoon keinotekoista painotusta monistettujen alkioiden tapauksiin. Menetelmä saattaa tehdä harvinaisesta tapauksesta yleisen.

Kubat, Holte ja Matwin mainitsevat, että neuroverkkojen yliopetus voi olla vaarallista. Näin menettelemällä voitaisiin kuitenkin varmistua, että järjestelmä oppii myös pienemmän luokan ominaisuudet. Yliopetus voi tapahtua käyttämällä tehokasta verkkoa tai suorittamalla ylimääräisiä opetusiteraatioita. Ylioppimista kuitenkin tulisi välttää, sillä se heikentää verkon kykyä yleistää [11].

Keinotekoista dataa on vaarallista käyttää sekvenssien yhteydessä, koska luokat menevät pahasti toistensa päälle. Muutoinkin keinotekoisen datan käyttöön näyttäisi liittyvän periaatteellinen ongelma. Neuroverkon pitäisi oppia ilmiön synnyttämästä datasta oleelliset piirteet. Jos aineistoon generoidaan ylimääräisiä ennestään tuntemattomia syötekombinaatioita, kuinka voidaan olla varma mitä ilmiötä tapaukset edustavat?

Ainoa luonteva keino epätasapainon poistamiseen on dominoivan luokan pienentäminen. Pienentämistä käytettiin tässäkin työssä. Luokan ei-PPII-aineistoa karsittiin osaksi systemaattisella otannalla ja osaksi satunnaisotannalla.

Jos luokat esiintyvät luonnossa voimakkaasti epätasapainossa, ei neuroverkon ennustustarkkuuskaan ole aivan yksiselitteinen asia. Kubat ja Matwin

ovat pohtineet ennustustarkkuutta uudesta näkökulmasta [14].

Merkitään luvulla a oikeinmenneiden ei-PPII-ennustusten lkm., luvulla b väärinmenneiden ei-PPII-ennusteiden lkm., luvulla c väärinmenneiden PPII-ennusteiden lkm. ja luvulla d ne tapaukset, jossa PPII-tapaus on ennustettu oikein.

Perinteinen ennustustarkkuus määritellään oikeinmenneiden ennusteiden suhteena kaikkiin ennusteisiin eli

$$ennustustarkkuus = \frac{a+d}{a+b+c+d}. \tag{17}$$

Kubat ja Matwin suosittelevat edellisen asemasta käytettäväksi geometrista keskiarvoa eli

$$ennustustarkkuus = \sqrt{\frac{a}{a+b} * \frac{d}{c+d}}. \tag{18}$$

Ennustustarkkuudesta tulee hyvä vain, jos kummatkin ennusteet ovat hyviä [15].

Toinen aineiston vakava ongelma on päällekkäiset luokat. Frekvenssitaulukosta 2 nähdään, että ei ole olemassa sellaisia aminohappoja, jotka esiintymisellään ratkaisisivat kumpaan luokkaan tapaus kuuluu. Parhaimmissa tapauksissa aminohapot G, H, L, N, P, S, V ja Y antavat viitteitä siitä, suurentavatko vai pienentävätkö ne rakenteen esiintymistodennäköisyyttä.

Luokat ovat muuttujien suhteen voimakkaasti päällekkäin. Erotteleva tekijä voi kuitenkin olla yksittäisten muuttujien kombinaatiot. Neuroverkkojen pitäisi pystyä löytämään nämä vuorovaikutuksessa olevat muuttujat.

5 Tulokset

Työssä haluttiin selvittää, voiko PPII-rakenteen esiintymistä ennustaa primaarirakenteen perusteella. Kysymys on siitä, onko opetusjoukossa riittävästi sellaista informaatiota, jonka avulla testijoukko osataan luokitella. Tämä ratkaistaan tekemällä opetus ja testaus useita kertoja useilla opetusryhmillä.

Aluksi tarkastellaan neuroverkon tekemiä ennusteita tasanjakautuneen testijoukon suhteen. Pääpaino on sopivien menetelmien löytymisessä. Seuraavaksi tarkastellaan tuloksia luonnollisen jakauman tapauksessa ja kokeillaan keinoja parantaa tuloksia. Lopuksi esitellään, millaisia ovat ja missä sijaitsevat ne tapaukset, joita neuroverkko ei osaa luokitella oikein.

5.1 Tasanjakautunut testiaineisto

Tarkastellaan kysymystä, millainen olisi verkon ennustuskyky, jos PPII-rakennetta esiintyisi yhtä paljon kuin kaikkia muita rakenteita yhteensä. Tällä tavoin saatu ennustustarkkuus on keinotekoinen, mutta siitä pystyy näkemään ilmiön yleisluonteen. Tasanjakautuneen testiaineiston avulla etsitään myös paras tarkasteluikkunan pituus, ikkunointimenetelmä sekä piilosolmujen lukumäärä.

5.1.1 Neljä opetusryhmää

Aluksi neuroverkko opetettiin neljällä eri opetusryhmällä, kahdella eri ikkunointiratkaisulla (1 ja 2), kahdella sekvenssin pituudella (7 ja 13) sekä neljällä eri piilosolmun lukumäärällä.

Taulukoissa olevat alitaulukot (edellisen luvun lopussa on esimerkki tällaisesta) kuvaavat aina yhden opetetun neuroverkon antamaa ennustetulosta testijoukolle. Yhdeksi ryhmäksi (erotuksena opetusryhmästä) kutsutaan yhtä neljän verkon kokonaisuutta, jolla on yhteinen testijoukko. Ryhmät jakavat taulukon vasempaan ja oikeaan sarakkeeseen. Yhdessä ryhmässä neuroverkkojen tulokset esitetään yhdellä tarkasteluikkunan pituudella ja tietyllä ikkunointimenetelmällä. Ryhmää käsitellään yhtenä kokonaisuutena ja vain huomionarvoiset yksityiskohdat poimitaan. Tuloksissa esitellään

Taulukko 4. Neuroverkon antamat ennusteet opetusryhmän 1 ikkunointimenetelmälle 1 ja 2. Tarkasteluikkunan koko on 13 ja piilosolmujen lukumäärä 4 - 25.

testijoukossa 1250 kpl.

1, 13, 4		ennustus: negat	ennustus: posit
todellisuus	negat	37.6	12.4
	posit	16.7	33.3

1, 13, 8		negat	posit
	negat	31.1	18.9
	posit	11.2	38.8

1, 13, 15		negat	posit
	negat	36.3	13.7
	posit	15.7	34.3

1, 13, 25		negat	posit
	negat	38.8	11.2
	posit	17.1	32.9

testijoukossa 1980 kpl.

2, 13, 4		ennustus: negat	ennustus: posit
todellisuus	negat	39.4	10.6
	posit	20.2	29.8

2, 13, 8		negat	posit
	negat	37.6	12.4
	posit	19.1	30.9

2, 13, 15		negat	posit
	negat	34.6	14.9
	posit	14.7	35.8

2, 13, 25		negat	posit
	negat	36.9	13.1
	posit	17.3	32.7

kummankin luokan tapausten löytymis- ja ennustustarkkuudet sekä näiden yhteinen ennustustarkkuus.

Yhdestä neuroverkosta käytetään tunnistetta *verkko(e,f,g)*, jossa symboli e on ikkunointimenetelmä, f on tarkasteluikkunan koko ja g on piilosolmujen lukumäärä. Tunniste on alitaulukon vasemmassa yläreunassa.

Taulukon 4 tulokset on saatu opetusryhmälle 1. Vasemmalla puolella ovat neuroverkon ennusteet ikkunointimenetelmällä 1 ja tarkasteluikkunan pituudella 13. Piilosolmuja on käytetty 4 - 25 kpl. Opetusjoukossa oli yhteensä 8882 tapausta ja testijoukossa yhteensä 1250 tapausta.

Ryhmässä on poikkeustapauksena verkko (1,13,25). Verkko sisältää 25 piilosolmua ja verkon yhteyksien määrä on yli 7500 kpl. Sääntöjen mukaan tämä vaatisi opetusaineistoa 75000 tapausta, mutta todellisuudessa aineistoa oli käytettävissä vajaat 9000. Verkko löytää ryhmänsä huonoiten PPII-rakenteita.

Keskimäärin ryhmä löytää testiaineistosta 69.6 % PPII-tapauksia ja 71.9 % ei-PPII-tapauksia. Ennustustarkkuus PPII-rakenteille on 71.6 % ja ei-PPII-rakenteille 70.5 %. Ryhmän kokonaisennustetarkkuus on 70.8 %. Verkon (1,13,8) tapauksessa PPII-tapauksia on poikkeuksellisesti löydetty enemmän kuin ei-PPII-rakenteita. Ilmiön tekee mielenkiintoiseksi se, että samaisella verkolla ei-PPII-tapauksia löydetään heikommin.

Oikeanpuoleiset tulokset on saatu ikkunointimenetelmällä 2. Tarkastelu-

Taulukko 5. Neuroverkon antamat tulokset opetusryhmälle 2. Käytössä on ollut ikkunointimenetelmä 1. Mukana on tarkasteluikkunan pituudet 12 ja 7 ja piilosolmujen lukumäärät 2 - 15.

testijoukossa 1366 kpl.

1, 13, 2		ennustus: negat	ennustus: posit
todellisuus	negat	37.6	12.4
	posit	11.0	39.0

1, 13, 4		ennustus: negat	ennustus: posit
todellisuus	negat	36.7	13.3
	posit	10.3	39.7

1, 13, 8		ennustus: negat	ennustus: posit
todellisuus	negat	36.6	13.4
	posit	11.5	38.5

1, 13, 15		ennustus: negat	ennustus: posit
todellisuus	negat	36.5	13.5
	posit	10.9	39.1

testijoukossa 1404 kpl.

1, 7, 2		ennustus: negat	ennustus: posit
todellisuus	negat	38.0	12.0
	posit	12.7	37.3

1, 7, 4		ennustus: negat	ennustus: posit
todellisuus	negat	37.6	12.4
	posit	13.0	37.0

1, 7, 8		ennustus: negat	ennustus: posit
todellisuus	negat	37.6	12.4
	posit	12.7	37.3

1, 7, 15		ennustus: negat	ennustus: posit
todellisuus	negat	38.1	11.9
	posit	12.5	37.5

ikkunan pituus on ollut 13 ja piilosolmujen lukumäärät ovat olleet välillä 4 - 25. Opetusjoukossa oli 14138 tapausta ja testijoukossa tapauksia oli 1980. PPII-tapauksia löydetään keskimäärin 64.4 % ja ei-PPII-tapauksia 74.4 %. Ennustustarkkuudeksi saadaan 71.7 % PPII-tapauksille ja 67.7 % ei-PPII-tapauksille. Kokonaisennustustarkkuus on 69.4 %.

Tässä ryhmässä esiintyy toinen verkko, jossa on 25 piilosolmua (verkko (2,13,25). Vaikka opetustapauksia on aikaisempaa ryhmää enemmän, ei niitä ole läheskään riittävästi näin suurelle verkolle. Verkko ei osoita suurempia kykyjä PPII-tapauksiin nähden ja kokonaisennustustarkkuus on 69.5 %. Heikkojen ennusteiden ja suurten yhteyksien määrän takia yli viidentoista piilosolmun verkkoja ei opeteta enää. Mukaan otetaan verkko, jossa on 2 piilosolmua.

Kokonaisennustetarkkuudet ovat vasemmanpuoleisissa verkoissa paremmat - ikkunointimenetelmä 1 on ollut tehokkaampi. Lisäksi kannattaa huomioida luokkien saamien ennusteiden negatiivinen korrelointi. Kun toinen luokka saa hyviä arvoja, toisen luokan suhteen tulos on huonompi. Näin ei tarvitsisi välttämättä olla, sillä verkon pitäisi pystyä riippumattomasti lajittelemaan PPII- ja ei-PPII-tapaukset.

Taulukossa 5 on opetusryhmällä 2 opetettujen verkkojen tulokset. Käytössä on ollut ikkunointimenetelmä 1 ja tarkasteluikkunan pituudet 13 ja 7. Kummassakin ryhmässä on nyt mukana verkot, joissa on 2 piilosolmua.

Taulukko 6. Neuroverkon antamat tulosteet opetusryhmän 2 ikkunointiratkaisulle 2. Mukana tarkasteluikkunan pituudet 13 ja 7. Piilosolmujen lukumäärä vaihtelee välillä 2 - 15.

testijoukossa 2188 kpl.

2, 13, 2		ennustus: negat	ennustus: posit
todellisuus	negat	36.6	13.4
	posit	12.3	37.2

2, 13, 4		ennustus: negat	ennustus: posit
	negat	36.4	13.6
	posit	13.1	36.9

2, 13, 8		ennustus: negat	ennustus: posit
	negat	36.1	13.9
	posit	11.9	38.1

2, 13, 15		ennustus: negat	ennustus: posit
	negat	36.1	13.9
	posit	12.9	37.1

testijoukossa 2242 kpl.

2, 7, 2		ennustus: negat	ennustus: posit
todellisuus	negat	37.0	13.0
	posit	14.9	35.1

2, 7, 4		ennustus: negat	ennustus: posit
	negat	36.8	13.2
	posit	15.2	34.8

2, 7, 8		ennustus: negat	ennustus: posit
	negat	36.8	13.2
	posit	14.4	35.6

2, 7, 15		ennustus: negat	ennustus: posit
	negat	36.1	13.9
	posit	14.5	35.5

Vasemmanpuoleiset verkot on opetettu 13-mittaisella aineistolla. Aineistossa on ollut mukana 8864 opetustapausta ja 1366 testitapausta. Testijoukosta löydetään 77.6 % PPII-rakenteita ja 74.2 % ei-PPII-rakenteita. Tällöin PPII-luokan ennustustarkkuus on 75.3 % ja ei-PPII-luokan 76.6 %. Ryhmän kokonaisennustustarkkuus on 75.9 %.

Oikeanpuoleinen ryhmä on opetettu aineistolla, jossa tarkasteluikkunan pituus on ollut 7 aminohappoa. Aineistossa on ollut 8678 opetustapausta ja 1404 testitapausta. PPII-tapauksia on kokonaismäärästä löydetty 74.5 % ja ei-PPII-tapauksia 75.6 %. Ennustustarkkuus PPII-luokan suhteen on 75.4 % ja ei-PPII-luokan suhteen 74.8 %. Kokonaisennustustarkkuus on myös varsin korkea eli 75.1 %.

Jokainen vasemman puoliskon verkko ennusti paremmin PPII-tapauksia kuin ei-PPII-tapauksia. Edellisessä opetusryhmässä tilanne oli päinvastainen. Oikeanpuolen tulokset ovat suunnilleen tasapainossa luokkien suhteen. Kokonaisuudessaan ennustearvot ovat parhaat juuri opetusryhmässä 2. Ehkä ryhmään on sattunut etsiytymään paljon opetusjoukon kaltaisia tapauksia.

Opetusryhmälle 2 saatuja tuloksia esitellään myös taulukossa 6. Käytössä on ollut ikkunointimenetelmä 2 ja tarkasteluikkunan pituudet 13 ja 7.

Vasen ryhmä on opetettu aineistolla, joka on saatu 13-mittaisella tarkasteluikkunalla. Aineistossa on ollut mukana 14088 opetustapausta ja 2188 testitapausta. Verkot luokittelevat oikein keskimäärin 74.9 %:n tarkkuudella PPII-tapaukset ja 72.5 %:n tarkkuudella ei-PPII-tapaukset. Ennustustark-

kuus PPII-luokan rakenteille on 73.2 % ja ei-PPII-luokan rakenteille 74.3 %. Kokonaisennustustarkkuus on keskimäärin 73.7 %.

Oikeanpuoleinen ryhmä on opetettu aineistolla, jossa tarkasteluikkunan pituus on ollut 7. Aineistossa on 13708 opetustapausta ja 2242 testitapausta. PPII-tapauksista löydetään 70.5 % ja ei-PPII-tapauksista löydetään 73.3 %. Ennustustarkkuus PPII-luokassa on 72.6 % ja ei-PPII-luokassa 71.3 %. Kokonaisennustustarkkuus on 72.0 %.

Taulukoiden 4, 5 ja 6 mukaan ikkunointimenetelmä 1 antaa parempia tuloksia. Taulukoista nähdään vielä, että opetusryhmällä 2 saadaan parempia tuoksia kuin opetusryhmällä 1.

Taulukossa 7 on ensimmäinen osa tuloksista opetusryhmälle 3. Käytössä on ollut ikkunointiratkaisu 1 ja tarkasteluikkunassa on käytetty pituuksia 13 ja 7.

Vasemmalla puolella oleva ryhmä on opetettu 13-mittaisella aineistolla. Aineistossa on ollut 8676 opetussekvenssiä ja 1614 testisekvenssiä. PPII-tapauksista löydetään 71.9 % ja ei-PPII-tapauksista 74.8 %. Ennustustarkkuudeksi saadaan tällöin PPII-rakenteille 74.1 % ja ei-PPII-rakenteille 72.7 %. Ryhmän kokonaisennustustarkkuus on 73.4 %.

Oikeanpuoleinen ryhmä on opetettu aineistolla, jossa tarkasteluikkunan pituus on ollut 7. Aineistossa on 8500 opetustapausta ja 1648 testitapausta. Kokonaismäärästä on löydetty 73.0 % PPII-rakenteita ja 75.1 % ei-PPII-rakenteita. Ennustustarkkuus PPII-luokassa on 74.1 % ja ei-PPII-luokassa 72.7 %. Kokonaisennustustarkkuus on 74.0 %.

Kummassakin taulukon 7 ryhmässä ei-PPII-rakenteita löydetään paremmin.

Taulukon 8 tulokset ovat myös opetusryhmän 3 aineistolle. Aineiston hankintaan on käytetty ikkunointiratkaisua 2. Tarkasteluikkunan pituudet ovat 13 ja 7.

Tarkasteluikkunan pituudella 13 on saatu opetusaineistoa 13810 tapausta ja testiaineistoa 2540 tapausta. PPII-rakenteita löydetään 71.8 % ja ei-PPII-rakenteita 73.8 %. PPII-rakenteille ennustustarkkuudeksi saadaan 73.2 % ja ei-PPII-rakenteille 72.3 %. Ryhmän kokonaisennustustarkkuus on 72.8 %.

Oikeanpuoleinen ryhmä on opetettu aineistolla, jossa tarkasteluikkunan pituus on ollut 7. Aineistossa on 13468 opetussekvenssiä ja 2582 testisekvenssiä. Kokonaismäärästä on löydetty vain 68.6 % PPII-rakenteita ja ei-PPII-rakenteita 74.3 %. Ennustustarkkuus PPII-luokassa on 72.8 % ja ei-PPII-luokassa 70.3 %. Kokonaisennustustarkkuus on 71.5 %.

Näissäkin tuloksissa piilosolmujen lukumäärä 4 antaa tasapainoisimman tuloksen luokkien välille. Lyhyemmällä sekvenssillä opetetut ennustavat hie-

Taulukko 7. Neuroverkon antamia tuloksia opetusryhmän 3 ikkunointiratkaisulle 1. Tarkasteluikkunan pituudet ovat 13 ja 7 ja piilosolmujen lukumäärät vaihtelevat välillä 2 - 15.

testijoukossa 1614 kpl.

1, 13, 2		ennustus: negat	ennustus: posit
todellisuus	negat	37.5	12.5
	posit	14.2	35.8

1, 13, 4	ennustus: negat	ennustus: posit
negat	37.5	12.5
posit	13.9	36.1

1, 13, 8	ennustus: negat	ennustus: posit
negat	37.1	12.9
posit	13.6	36.4

1, 13, 15	ennustus: negat	ennustus: posit
negat	37.5	12.5
posit	14.4	35.6

testijoukossa 1648 kpl.

1, 7, 2		ennustus: negat	ennustus: posit
todellisuus	negat	37.0	13.0
	posit	13.1	36.9

1, 7, 4	ennustus: negat	ennustus: posit
negat	37.7	12.3
posit	13.3	36.7

1, 7, 8	ennustus: negat	ennustus: posit
negat	37.9	12.1
posit	13.5	36.5

1, 7, 15	ennustus: negat	ennustus: posit
negat	37.5	12.5
posit	14.1	35.9

Taulukko 8. Neuroverkon antamat tulokset opetusryhmän 3 ikkunointiratkaisuille 2. Mukana tarkasteluikkunan pituudet 13 ja 7. Piilosolmujen lukumäärät ovat välillä 2 - 15.

testijoukossa 2540 kpl.

2, 13, 2		ennustus: negat	ennustus: posit
todellisuus	negat	36.9	13.1
	posit	14.4	35.6

2, 13, 4	ennustus: negat	ennustus: posit
negat	36.5	13.5
posit	13.4	36.6

2, 13, 8	ennustus: negat	ennustus: posit
negat	37.0	13.0
posit	14.6	35.4

2, 13, 15	ennustus: negat	ennustus: posit
negat	37.1	12.9
posit	14.1	35.9

testijoukossa 2582 kpl.

2, 7, 2		ennustus: negat	ennustus: posit
todellisuus	negat	37.3	12.7
	posit	15.6	34.4

2, 7, 4	ennustus: negat	ennustus: posit
negat	37.1	12.9
posit	15.4	34.6

2, 7, 8	ennustus: negat	ennustus: posit
negat	37.1	12.9
posit	16.1	33.9

2, 7, 15	ennustus: negat	ennustus: posit
negat	37.2	12.8
posit	15.6	34.4

man voimakkaammin ei-PPII-tapauksia. Ehkä PPII-rakenteen informaatio sijaitsee etäämmällä itse rakenteesta.

Taulukossa 9 on tulokset opetusryhmän 4 aineistolle. Aineiston hankintaan on käytetty ikkunointiratkaisua 1. Tarkasteluikkunan pituudet ovat 13 ja 7.

Vasemmanpuoleisten verkkojen opetusaineistossa on ollut 8782 tapausta ja testijoukossa 1362 tapausta. Aineisto on saatu ikkunointimenetelmällä 1 ja tarkasteluikkunan pituuksilla 13 ja 7. PPII-tapauksista löydetään 66.4 % ja ei-PPII-tapauksista 74.4 %. Ennustustarkkuudeksi saadaan tällöin PPII-rakenteille 72.2 % ja ei-PPII-rakenteille 68.9 %. Ryhmän kokonaisennustustarkkuus on 70.4 %.

Oikeanpuoleinen ryhmä on opetettu aineistolla, jossa tarkasteluikkunan pituus on ollut 7. Aineistossa on 8596 opetustapausta ja 1404 testitapausta. Kokonaismäärästä on löydetty 66.0 % PPII-rakenteita ja ei-PPII-rakenteita 75.5 %. Ennustustarkkuus PPII-luokassa on 72.9 % ja ei-PPII-luokassa 68.9 %. Kokonaisennustustarkkuus on 70.7 %.

Näillä verkoilla löydetään PPII-rakenteita todella huonosti. Syynä ei voi olla epätasainen neuroverkon alustuskaan, sillä sama ilmiö toistuu jokaisessa kahdeksassa verkossa.

Taulukko 9. Neuroverkon antamat ennusteet ikkunointimenetelmälle 1. Tarkasteluikkunan pituudet ovat 13 ja 7. Piilosolmujen lukumäärät vaihtelevat välillä 2 - 15.

testijoukossa 1362 kpl.

1, 13, 2		ennustus: negat	ennustus: posit
todellisuus	negat	36.5	13.5
	posit	16.2	33.8

1, 13, 4	negat	posit
negat	37.5	12.5
posit	16.8	33.2

1, 13, 8	negat	posit
negat	37.4	12.6
posit	16.5	33.5

1, 13, 15	negat	posit
negat	37.4	12.6
posit	17.5	32.5

testijoukossa 1404 kpl.

1, 7, 2		ennustus: negat	ennustus: posit
todellisuus	negat	38.0	12.0
	posit	17.3	32.7

1, 7, 4	negat	posit
negat	37.6	12.4
posit	12.1	32.9

1, 7, 8	negat	posit
negat	37.4	12.6
posit	17.2	32.8

1, 7, 15	negat	posit
negat	38.0	12.0
posit	16.5	33.5

Taulukko 10. Neuroverkon antamat ennusteet ikkunointimenetelmälle 2. Tarkasteluikkunan pituudet ovat 13 ja 7. Solmujen lukumäärä vaihtelee välillä 2 - 15.

testijoukossa 2139 kpl.

2, 13, 2		ennustus: negat	ennustus: posit
todellisuus	negat	36.2	13.8
	posit	18.3	31.7

2, 13, 4	negat	posit
negat	35.7	14.3
posit	17.5	32.5

2, 13, 8	negat	posit
negat	35.0	15.0
posit	17.7	32.3

2, 13, 15	negat	posit
negat	36.5	13.5
posit	18.6	31.4

testijoukossa 2210 kpl.

2, 7, 2		ennustus: negat	ennustus: posit
todellisuus	negat	36.7	13.3
	posit	19.1	30.9

2, 7, 4	negat	posit
negat	36.7	13.3
posit	19.5	30.7

2, 7, 8	negat	posit
negat	36.6	13.4
posit	19.2	30.8

2, 7, 15	negat	posit
negat	36.5	13.5
posit	19.2	30.8

Taulukon 10 tulokset on saatu myös opetusryhmän 4 aineistolle. Ikkunointiratkaisuna on käytetty menetelmää 2 ja tarkasteluikkunan pituudet ovat 13 ja 7.

Vasemmanpuoleisten verkkojen opetusaineistossa on ollut 13986 tapausta ja testijoukossa 2138 tapausta. PPII-tapauksista löydetään 63.9 % ja ei-PPII-tapauksista 71.7 %. Ennustustarkkuudeksi saadaan tällöin 69.3 % PPII-rakenteille ja 66.5 % ei-PPII-rakenteille. Ryhmän kokonaisennustustarkkuus on 67.8 %.

Oikeanpuoleinen ryhmä on opetettu aineistolla, jossa tarkasteluikkunan pituus on ollut 7. Aineistossa on 13628 opetustapausta ja 2210 testitapausta. Kokonaismäärästä on löydetty ainoastaan 61.5 % PPII-rakenteita ja 73.3 % ei-PPII-rakenteita. Ennustustarkkuus PPII-luokassa on 69.8 % ja ei-PPII-luokassa 65.6 %. Kokonaisennustustarkkuus on 67.4 %.

Nämäkin verkot löytävät todella huonosti PPII-luokan tapaukset. PPII-luokkaan on sattunut juuri sellaisia tapauksia, jotka ovat vaikeasti erotettavissa ei-PPII-luokasta. Taulukossa 9 ja 10 on yhteisenä ominaisuutena sama opetusryhmä, joten ikkunointimenetelmä 1 näyttää olleen tehokkaampi.

5.1.2 Parhaimpien menetelmien valinta

Edellä esitellyissä opetusryhmissä on yhteensä 28 ikkunointimenetelmällä 1 opetettua neuroverkkoa ja 28 ikkunointimenetelmällä 2 opetettua neuroverkkoa. Tarkasteluikkunan pituudella 13 on opetettu 32 verkkoa ja tarkasteluikkunan koolla 7 on opetettu 24 verkkoa. Kahta piilosolmua on käytetty 12:ssa neuroverkossa, neljää, kahdeksaa ja viittätoista piilosolmua 14:ssä verkossa. Suuria 25:n piilosolmun neuroverkkoja opetettiin vain kaksi kappaletta.

Ensimmäiseksi testataan pituuden merkityksestä ennustamistulokseen. Testaus tehdään merkkitestillä, joka perustuu binomitodennäköisyyksien avulla määriteltyyn kriittiseen alueeseen. Tällöin järjestysasteikollinen muuttuja riittää. Kun otoksessa on enemmän tapauksia kuin 20, voidaan tehdä normaalijakauma-aproksimaatio. [9]

Testauksessa toisen ryhmän muodostavat 5- ja 7-mittaisella aineistolla opetetut verkot (ryhmä 5-7) ja toisen ryhmän muodostavat 13-mittaisella aineistolla opetetut verkot (ryhmä 13).

Kun tulokset asetetaan rinnakkain, saadaan 21 kpl. ryhmän 13 verkkoa, jotka antavat paremman tuloksen kuin ryhmän 5-7 verkot. Vastaavasti löytyy 11 ryhmän 5-7 verkkoa, jotka antavat paremman tuloksen kuin ryhmän 13 verkot.

Muodostetaan nollahypoteesi: pituudella ei ole merkitystä ennustustarkkuuteen. Tämän vastahypoteesi on, että toinen menetelmistä antaa paremman tuloksen kuin toinen.

Asetetaan luottamusrajaksi $\alpha = 0.05$, jolloin normaalijakaumataulukosta nähdään, että testisuureen Z pitää olla suurempi tai yhtäsuuri kuin 1.65. Testisuure Z lasketaan kaavalla

$$\frac{\check{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}. \tag{19}$$

Kaavassa $\check{p}$ on suuremman lukumäärän suhde kaikkien tapausten lukumäärään eli tässä tapauksessa $21/(21+11)$. Symboli P_0 on nollahypoteesin mukainen osuus luokille eli 0.5 ja n on havaintojen lukumäärä 32. Testisuureelle Z saadaan arvo 1,76, joten nollahypoteesi voidaan hylätä riskillä 0.05 ja testin p-arvoksi saadaan 0.04. Pidemmän tarkasteluikkunan käytto antaa 95 %:n todennäköisyydellä paremman tuloksen.

Vastaava testaus tehtiin ikkunointiratkaisun 1 ja 2 välillä. Merkkitesti antoi testisuureelle Z arvon 4.59. Nollahypoteesi (ikkunointimenetelmällä ei ole merkitystä ennustustarkkuuteen) voidaan hylätä riskillä 0.05 ja p-arvo oli huikeat 0.00001. Ikkunointimenetelmää 1 voidaan siis varmasti pitää parempana.

Viimeinen testaus tehtiin piilosolmujen lukumäärän suhteen. Testissä muodostettiin kaksi ryhmää: neuroverkot, joissa oli 2 tai 4 piilosolmua ja neuroverkot, joissa oli 8, 15 ja 25 piilosolmua. Nollahypoteesi oli, että piilosolmujen lukumäärällä ei ole merkitystä ennustustarkkuuteen. Vaihtoehtoinen hypoteesi oli päinvastainen.

Testisuureelle Z saatiin arvo 1.89, joten piilosolmujen lukumäärällä on vaikutusta ennustustarkkuuteen. Nollahypoteesi voidaan hylätä riskillä 0.05 ja testin p-arvo on 0.03. Siis neljän tai alle neljän piilosolmun verkoilla saadaan paremmat ennustetulokset kuin suuremmilla verkoilla. Tämä on hieman yllättävää, sillä 13 aminohapon pituisessa syötteessä syötekerrokselta saapuu jokaiselle piilokerroksen solmulle 299 yhteyttä.

5.1.3 Keskiarvot kahdeksasta opetusryhmästä

Kun tiedetään, millainen verkon ja aineiston tulee olla, voidaan pyrkiä kohti luotettavaa keskiarvoa. Tässä vaiheessa opetettiin kahdeksalla eri opetusryhmällä 24 neuroverkkoa. Sekvensseissä käytettiin pituutta 13, ikkunoinnissa käytettiin menetelmää 1 ja piilosolmujen lukumäärää 4. Tästä saatiin taulukon 11 ennustetulokset.

Taulukko 11. Kahdeksalle eri opetusryhmälle saadut ennustetulokset. Yksi alitaulukko kuvaa kokonaisennusteeltaan parasta kolmesta eri alustuksesta.

testijoukossa 1250 kpl.

1	ennustus: negat	ennustus: posit
todellisuus: negat	37.4	12.6
todellisuus: posit	15.4	34.6

testijoukossa 1780 kpl.

5	ennustus: negat	ennustus: posit
todellisuus: negat	37.2	12.8
todellisuus: posit	13.4	36.6

testijoukossa 1366 kpl.

2	negat	posit
negat	36.9	13.1
posit	9.8	40.2

testijoukossa 1734 kpl.

6	negat	posit
negat	35.9	14.1
posit	11.5	38.5

testijoukossa 1614 kpl.

3	negat	posit
negat	37.3	12.7
posit	13.3	36.7

testijoukossa 1676 kpl.

7	negat	posit
negat	38.8	11.2
posit	15.0	35.0

testijoukossa 1360 kpl.

4	negat	posit
negat	37.4	12.7
posit	16.5	33.4

testijoukossa 1528 kpl.

8	negat	posit
negat	37.8	12.2
posit	14.5	35.5

Neuroverkkojen antamista tuloksista voidaan laskea, että PPII-rakenteita pystytään löytämään keskimäärin 72.6 % kaikista PPII-tapauksista. Tämä luonnollisesti tarkoittaa sitä, että neuroverkot eivät kykene erottamaan loppuja ei-PPII-rakenteista. Neuroverkot pystyvät löytämään keskimäärin 74.7 % ei-PPII-rakenteista ja loppuja neuroverkot pitävät PPII-tapauksina.

Löydettyjen PPII-tapausten suhde kaikkiin PPII-tapauksiksi ennustettuihin on 0.741, joten ennustustarkkuus PPII-luokalle on 74.1 %. Vastaavasti ei-PPII-tapausten suhde kaikkiin ei-PPII-tapauksiksi ennustettuihin on 0.733, joten ennustustarkkuus ei-PPII-luokalle on 73.3 %. Kokonaisuudessaan ennustustarkkuus on kaikkien oikeinluokiteltujen suhde testijoukon tapausten lukumäärään eli 0.737 ja ennustustarkkuus on tällöin 73.7 %.

5.1.4 Opetusaineisto, jossa ei ole identtisiä sekvenssejä

Yhdessä satunnaisesti valitusta opetusryhmästä poistettiin vastakkaisissa luokissa olevat identtiset tapaukset (opetusaineiston osalta). Verkko opetettiin kolmella eri alustuskerralla. Kahden parhaan verkon antamat tulokset ovat taulukossa 12. Vastaavalla aineistolla on opetettu myös taulukon 11 verkko 7.

Verkot löytävät hyvin ei-PPII-tapauksia, mutta PPII-tapauksille saadaan hieman huonompia tuloksia kuin parhaissa aikaisemmissa verkoissa. Ennustetulos PPII-luokalle paranee sen takia, kun ei-PPII-tapaukset pystytään tehokkaasti erottamaan PPII-rakenteista. Tällöin ennustustulos PPII-rakenteelle

Taulukko 12. Tulokset opetusryhmästä, josta oli poistettu myös vastakkaisten luokkien identtiset alkiot.

testijoukossa 1362 kpl.

	1	ennustus: negat	ennustus: posit
todellisuus	negat	37.6	12.4
	posit	16.4	33.6

	5	ennustus: negat	ennustus: posit
todellisuus	negat	38.3	11.7
	posit	16.6	33.4

Taulukko 13. Neuroverkon antamat ennusteet opetusjoukolle, joka on saatu PPII-rakenteen pituudella 2.

testijoukossa 3340 kpl.

	1, 13, 4	ennustus: negat	ennustus: posit
todellisuus	negat	35.3	14.7
	posit	13.7	36.3

on 73.9 % ja ei-PPII-rakenteille 71.7 %.

5.1.5 Tulokset rakenteen pituudella 2

Rakenteita ikkunoitaessa voidaan käyttää löyhempiä ehtoja rakenteen pituuden suhteen. Tällöin saadaan reilusti enemmän opetusaineistoa ja tällä voi olla merkitystä neuroverkon oppimiseen, yleistämiseen ja ennustamiseen. Näin on tehty yhdelle opetusaineistolle, jonka tulokset ovat taulukossa 13.

Vaikka opetus- ja testiaineistoon saatiin runsaasti materiaalia, eivät tulokset ole parantuneet. PPII-tapauksista löydetään 72.0 % ja ei-PPII-tapauksista vain 70.6 %. Tämä tarkoittaa sitä, että ennustetulokset ovat PPII-luokalle vain 71.1 % ja ei-PPII-luokalle 72.0 %.

5.2 Luonnollisesti jakautunut testiaineisto

Edellä on esitelty tuloksia, joissa testijoukkoa on muutettu luonnollisen jakauman suhteen. Oikeassa ennustustehtävässä verkko joutuu kohtaamaan aina vääristymättömiä testiaineistoja.

PPII-rakenteita esiintyy pituudesta ja säännöllisyyskriteereistä riippuen yhdestä kolmeen prosenttiin. Taulukossa 14 esitellään ennustetuloksia, joita on saatu luonnolliselle jakaumalle ja rakenteen pituudelle 3.

Huomattavaa tuloksissa on se, että vaikka testijoukossa on huomattavan vino jakauma, on rakenteiden löytymisprosentti kuitenkin lähes vastaa-

Taulukko 14. Neuroverkon antamat ennusteet luonnollisesti jakautuneelle testiaineistolle.

testijoukossa 27484 kpl

1		ennustus negat	ennustus posit
todellisuus	negat	75.0	23.8
	posit	0.4	0.8

2		ennustus negat	ennustus posit
todellisuus	negat	73.1	25.7
	posit	0.4	0.8

testijoukossa 32260 kpl

3		negat	posit
todellisuus	negat	73.5	25.3
	posit	0.3	0.9

4		negat	posit
todellisuus	negat	73.2	25.6
	posit	0.3	0.9

va tasanjakautuneen aineiston kanssa. Löytymisprosentti on kummassakin luokassa yli 70 eli PPII- ja ei-PPII-rakenteista luokittuu väärin vain noin neljännes.

PPII-luokasta löydetään keskimäärin 72.0 % tapauksista ja ei-PPII-luokasta 74.5 %. Tästä voidaan laskea, että PPII-luokan ennustustarkkuus on 3.3 % ja ei-PPII-luokan 99.5 %. Kokonaisennustustarkkuus on 73.3 %. Nyt voidaan laskea ennustustarkkuuden ja esiintymistiheyden välille suhdeluku

$$\frac{3.3}{1.26} = 2.63. \tag{20}$$

Luonnollisen tapauksen kohdalla ennustuksen tekee ongelmalliseksi ne testijoukon ei-PPII-luokan tapaukset, jotka muistuttavat PPII-tapauksia. Näitä on PPII-tapauksiin verrattuna huomattavasti enemmän. Tämän takia ennustetulokset jäävät huonoiksi, vaikka PPII-tapauksista löydetään reilusti yli 70 %.

5.2.1 Peräkkäiset verkot

Tarkastellaan menetelmää, jossa testiaineisto luokitellaan peräkkäisillä neuroverkoilla kahteen kertaan.

Ensimmäinen opetus tapahtuu normaalisti aiemman kuvauksen tapaan. Ensimmäisen neuroverkon avulla luokitellaan opetusaineisto normaalisti. PPII-luokkaan määräytyneet sisältävät nyt oikeita PPII-tapauksia ja PPII-tapausten kaltaisia ei-PPII-tapauksia. Tästä saadaan uusi opetusaineisto, johon asetetaan kaikki pienemmän luokan tapaukset (väärinluokittuneet ei-PPII-tapaukset)

ja saman verran suuremman luokan alkioita. Neuroverkon tulisi uudella opetuskerralla oppia huomaamaan luokkia erottelevat hienommat piirteet.

Toisessa vaiheessa opetetaan uusi verkko, jolloin verkkoa voidaan käyttää aiemman verkon ennusteiden korjaamiseen. Uudella verkolla luokitellaan uudelleen vain PPII-luokkaan ennustetut tapaukset.

Aineistona käytettiin parhaan tuloksen antanutta opetusryhmää 2. Ensimmäinen lajittelu tehtiin taulukon 11 verkolla 2. Tällöin uuteen opetusjoukkoon saatiin 1033 kummankin luokan tapausta. Testijoukkoon laitetaan kaikki PPII-luokkaan ennustetut tapaukset eli 719 kpl. Uudella aineistolla opetettiin toinen neuroverkko.

Toinen verkko pystyi erottelemaan PPII-luokasta ei-PPII-luokkaan 30 tapausta. Näistä tapauksista oli yhdeksän ei-PPII-luokkaan kuuluvaa ja 21 PPII-luokkaan kuuluvaa. Verkko ei siis löydä enempää luokkia erottelevia piirteitä ja ohjautuu opetusjoukon mukana harhaan.

Menetelmää testattiin myös luonnolliselle jakaumalle. Eräällä aineistolla saatiin PPII-luokalle ennustustarkkuus 2.9 %. Toisen verkon jälkeen ennustustarkkuus oli 5.8 %. Ennustus kohosi kaksinkertaiseksi, mutta löydettyjen PPII-tapausten määrä tippui 74 %:sta 47 %:iin.

Eri luokkiin kuuluvien tapausten välille voidaan muodostaa uusi erotteleva kuvaus. Tällöin ennustustarkkuus saadaan kasvamaan, mutta rakenteita löytyy vähemmän.

5.2.2 Millaisia rakenteita verkko ennustaa

Verkko saattaa löytää luonnollisessa jakaumassa jopa 75 % PPII-tapauksista, mutta ennustaa useasti väärin. Millaisia ovat nuo väärinennustetut rakenteet? Taulukossa 15 esitetään tulkintoja löydetyistä rakenteista. Tulkinnat on saatu tarkastelemalla inhimillisellä otteella DSSP-tiedostossa olevia α-, ϕ- ja ψ-kulmia.

Tarkat säännöt eivät näytä luonnossa pätevän. Rakenteita löytyy sieltä missä niitä ennustetaankin olevan, mutta ne eivät sijaitse tarkalleen oikeassa paikassa tai niissä on mukana epätarkkuutta. Pitäisikö nämä hyväksyä oikein menneiksi ennusteiksi? Kysymystä voisi tarkentaa muotoon, onko hieman epäsäännöllisellä rakenteella samat biokemialliset ominaisuudet kuin täydellisellä rakenteella?

Ennustuksen osuvuutta tulee myös miettiä uudelleen. Koska kysymyksessä on harvinainen rakenne, voidaanko luopua täydellisen osuvuuden kriteeristä? Voitaisiinko kahden aminohapon etäisyydelle osuttua ennustetta pitää osumana?

Taulukko 15. Neuroverkkojen ennustamien rakenteiden kuvailua rakenteen pituudella 3.

proteiini	Neuroverkon ennustamia PPII-rakenteita: ennustetun rakenteen sijainti	kuvaus ennustuksesta
2ovw	2 - 5	rakenne on lyhyt (n = 2)
	10 - 12	alfa-kulma ei täsmää ja rakenne lyhyt
404 ennustettua PPII-rakennetta	25	hieman epäsäännöllinen fii-kulma
9 oikein	55 - 57	hieman sivussa (45 - 56)
395 väärin	62 - 67	alfa - kulma positiivinen (kierre oikeankätinen)
	71	rakenne lyhyt (n = 2) ja hieman sivussa (72 - 73)
	93	rajaa pienempi alfa-kulma (beta-tikapuu-rakenne)
	102 - 111	alueella lyhyt PPII-rakenne (n = 2)
	121	vieressä lyhyt PPII-rakenne (n = 2)
	128	ei rakenteeseen viittavia kulmia lähistöllä
	⋮	⋮
	852 - 854	oikein
	1251 - 1253	oikein
	⋮	⋮
1dpe	1 - 4	alueella lyhyt PPII-rakenne (n = 2)
	15 - 18	ei rakenteeseen viittavia kulmia lähistöllä
149 ennustettua PPII-rakennetta	25 - 31	ei rakenteeseen viittavia kulmia lähistöllä
1 oikein	45 - 48	alfa - kulma ei pysy rajoissa
148 väärin	66	oikein
	⋮	⋮

5.3 Spektrin antamaa tietoa

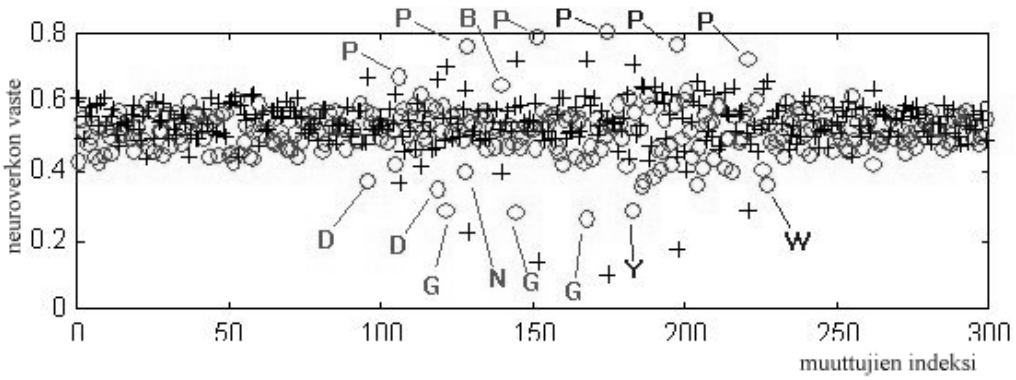
Neuroverkon "spektri" on havainnollistamismenetelmä, jolla nähdään, mitä neuroverkko on opetusaineistosta oppinut. Spektri paljastaa aineistosta samoja ilmiöitä kuin frekvenssi eli ilmiöön voimakkaasti vaikuttavia muuttujia. Spektriä esitellään tarkemmin liitteessä B.

Erään opetusryhmän antama spektri on kuvassa 17. Spektrikuvan vaaka-akseli on syötevektorin pituinen ja pystyakselilta näkee neuroverkon reagoinnin tiettyyn syötteeseen. Pallot kuvaavat PPII-luokkaa edustavan solmun tulostetta ja ristit ei-PPII-luokkaa edustavan solmun tulostetta. Kuviosta käsitellään vain PPII-luokkaa edustavan solmun vasteet, sillä toisen luokan solmun tulosteet hakeutuvat luvun 0.5 vastakkaiselle puolelle.

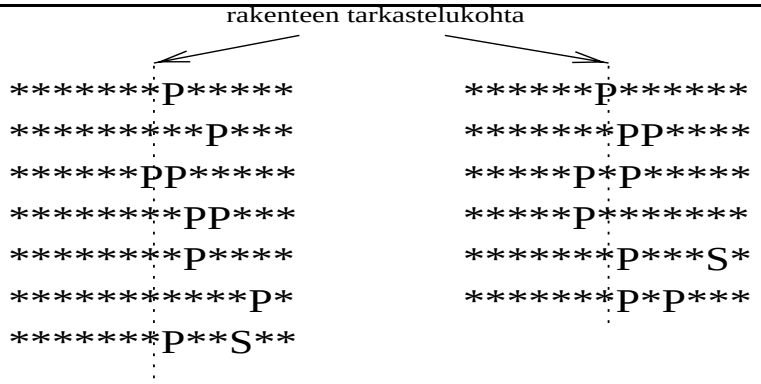
Neuroverkon tulosvektorissa on vain lukuja 0 ja 1. Tämän takia lukujen keskivaiheille (neutraalikohta) muodostuu tasainen neutraalien muuttujien "massa". Vaaka-akselin keskivaiheilla erottuu muutamia normaalia voimakkaampia muuttujia, koska PPII-rakenteen eiintymistä tarkastellaan tarkasteluikkunan keskimmäisen alkion kohdalta.

Voimakkaasti PPII-rakennetta edistävä aminohappo on odotetusti proliini P. Yllättävää puolestaan on, että aineistossa esiintyvä ylimääräinen kirjain B antaa neuroverkolle vihjeitä PPII-rakenteista.

Spektri paljastaa myös ne aminohapot, joiden lähistöllä PPII-rakennetta ei todennäköisesti esiinny. Näitä ovat frekvenssitaulukosta aikaisemmin löydetyt G, D, N, Y ja W.



Kuva 17: Neuroverkon antama tyypillinen spektrikuva opetusryhmän 3 aineistolle. Vaaka-akselilla on syötevektorin muuttujat ja pystyakselilla on neuroverkon tulossolmujen antamat tulosteet.



Kuva 18: Vasteanalyysin tuottamat PPII-luokalle edulliset aminohappokombinaatiot. Tähti tarkoittaa, että positioon ei tarvita mitään erityistä aminohappoa. Rakennetta tarkastellaan keskimmäisen aminohapon kohdalta.

5.4 Vasteanalyysin tulos

Vasteanalyysillä voidaan huomata usean muuttujan vuorovaikutussuhteet, vaikka nämä eivät frekvenssikuvaajissa näy. Tämä menetelmä antaa neuroverkolle syötekombinaatioita ja etsii näistä ilmiölle edullisia syötekombinaatioita. Myös tätä menetelmää esitellään liitteessä B. Vasteanalyysi tuotti PPII-luokalle kuvassa 18 esitetyt tulokset.

Tuloksissa näkyy proliinin P voimakas vaikutus rakenteen syntyyn. Yllättävää sen sijaan on, että neuroverkon mukaan proliineja ei tarvitse olla kolmea rakenteen todennäköisyyden kasvattamiseen.

Mielenkiintoinen on myös P:n vuorovaikutus aminohapon S kanssa. PPII-luokan tapauksissa esiintyi useita sekvenssejä, joissa aminohappo S oli sijoit-

tunut 2 - 5:n aminohapon välein. Osassa näistä ei tarvittu edes aminohappoa P.

Kun proteiineja ikkunoitiin, kiinnitettiin huomiota keskimmäisen aminohapon kohdalla sijaitsevaan rakenteeseen (ikkunointimenetelmässä 1). Kuvasta nähdään selvästi, että rakenteelle tärkeiden aminohappojen sijainti suhteessa rakenteeseen on keskimmäisen aminohapon jälkeisissä positioissa.

5.5 Hypoteesi väärinluokittuneista alkioista

Neuroverkko kykenee ennustamaan PPII- ja ei-PPII-luokan tapaukset hieman alle 75 prosentin tarkkuudella. Samalla tarkkuudella verkko pystyy luokittelemaan myös opetusjoukon alkiot. Tämä tarkoittaa sitä, että neuroverkko ei ole oppinut täydellisesti sekvenssien ja luokkien välistä kuvausta edes opetusjoukon suhteen.

Miksi verkko ei opi kuvausta paremmin? Tähän voidaan todeta, että ainakin testijoukon mukaan verkko on optimaalinen. Verkkoja kokeiltiin useilla eri piilosolmujen määrillä. Näistä eräät olivat jopa 30 piilosolmun verkkoja. Testijoukko saavutti optimin aina samoilla kohdilla. Neuroverkot eivät ehkä kykene luokittelemaan tapauksia, jotka sijaitsevat vastakkaisen luokan valta-alueella.

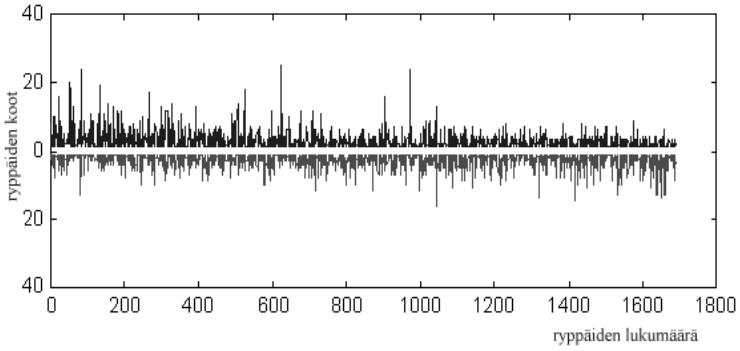
On mahdotonta tietää, onko todella näin. Ilmiötä voidaan kuitenkin yrittää tehdä helpommin ymmärrettäväksi. Liitteessä C esitellään menetelmä, jolla voidaan kuvata tapausten siroamista hahmoavaruuteen. Menetelmässä opetusaineisto luetellaan alueellisessa järjestyksessä ja pyritään löytämään yhtenäisiä alueita, joissa esiintyy vain toisen luokan alkioita.

Testattavaksi valittiin satunnaisesti yksi opetusryhmä, josta tutkittiin opetusjoukon erityyppisten tapausten siroamista. Tutkittavaksi valittiin opetusjoukko, koska neuroverkko rakentaa kuvauksen juuri tämän mukaan.

Aluksi opetusjoukolle tehtiin uusi luokitus, jossa jokaiselle alkiolle valittiin luokka arpomalla. Tällä yritetään antaa vertailukohta, millaisia sirontalukuja antaa täysin satunnaisesti hahmoavaruudessa olevat luokat.

Molempien syntyneiden ryhmien ryväskoko oli keskimäärin 1.98. Luokan 1 maksimiryväskoko oli 16 ja luokan 2 ryväskoko puolestaan 14. Ryhmän 1 ryväskokojen mediaani oli 2 ja ryhmän 2 vastaava luku oli 1. Menetelmän etsimiä vaihteluja syntyi 4466 kertaa, kun kaikkiaan alkioita oli 8864. Sirontasuhteeksi saadaan 0.50. Etsittäessä läheisintä alkiota on yhtä todennäköistä saada kummankin luokan tapaus. Tulos vaikuttaa järkevältä satunnaiselle sijoittelulle.

Toisessa ryhmässä analysoitiin koko opetusjoukko ja joukon alkioilla oli



Kuva 19: Sirontasuhteen mittauksen yhteydessä syntyneet ryväskoot koko opetusjoukolle. Nollakohdan yläpuolella on PPII-luokan ryppäät ja alapuolella ei-PPII-luokan ryppäät.

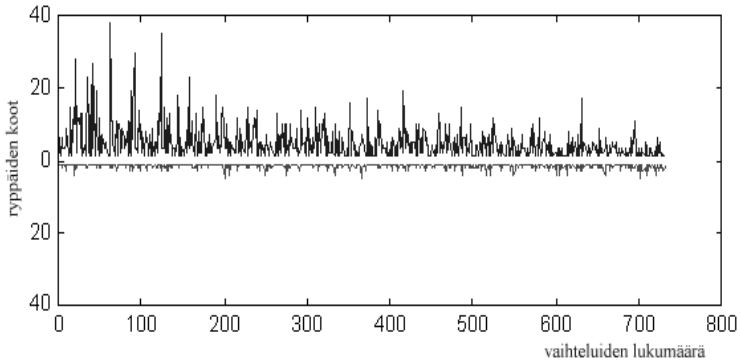
käytössä oikeat luokitukset. Toinen ryhmä antoi kuvan 19 mukaisen tuloksen. Nollakohdan yläpuolella on PPII-luokan ryppäät ja alapuolella on ei-PPII-luokan ryppäät.

PPII- sekä ei-PPII-luokassa keskimääräiseksi ryväskooksi saatiin 2.16. PPII-luokan ryppäiden maksimikoko oli 25 ja ei-PPII-luokalle vastaava koko oli 16. Kokonaisuudessaan vaihteluita luokkien välillä oli 3384. Kun tapauksia oli kaikkiaan 8864, saadaan sirontasuhteeksi 38.2 %. Tätä voidaan tietenkin pitää huonona, sillä etsittäessä läheistä saman luokan tapausta on melkein 40 %:n todennäköisyys törmätä vastakkaisen luokan tapaukseen.

Seuraavaksi opetusaineistosta poistettiin kaikki ne tapaukset, jotka luokittuvat vääriin luokkiin. Kuvasta rakentui lähes samanlainen kuin kuvassa 19. Ryväskoot ovat kasvaneet hieman. PPII-luokalle keskimääräinen ryväskoko oli 3.42 ja ei-PPII-luokalle 3.57. Maksimi ryväskoko PPII-luokalle oli 38 ja ei-PPII-luokalle 25. Tapauksia oli kaikkiaan 6654 ja luokkien välisiä vaihteluita esiintyi 1900, joten sirontasuhteeksi saadaan 28.5 %.

Kun analyysiin asetetaan PPII-luokkaan luokittuneet PPII-ja ei-PPII-tapaukset, saadaan kuvan 20 mukainen tulos. PPII-luokkaan kuuluvien alkioiden ryväskoko oli 4.44 ja ei-PPII-luokkaan kuuluvien 1.41. Maksimikoko PPII-luokan ryppäille oli 38 ja ei-PPII-luokan ryppäille 5. Luokkien välistä vaihtelua syntyi 1464 kappaletta ja vaihteluiden maksimimäärä on 2066 (kaksinkertainen määrä pienemmän luokan alkioita). Sirontasuhteeksi saadaan 70.8 %, joten pienemmän luokan tapaukset ovat todella sekaisin suuremman luokan seassa.

Viimeiseksi testiin asetettiin kaikki ne opetusjoukon tapaukset, jotka neuroverkko ohjaa ei-PPII-luokkaan. Kuvaaja on lähes vastaavanlainen kuin ku-



Kuva 20: Sirontasuhteen mittaamisen yhteydessä syntyneet ryväskoot PPII-luokkaan luokittuneille alkioille. Yläpuolella ovat PPII-luokan ryppäät ja alapuolella ovat ei-PPII-luokan ryppäät.

vassa 20. Luokan ei-PPII ryväskoko oli 4.50 ja luokan PPII ryväskoko oli 1.56. Maksimi ei-PPII-luokalle oli 32 ja vastustajaluokalle maksimi oli 7. Vaihtelua esiintyi 1509 kertaa ja kokonaisuudessaan vaihteluita pystyi muodostumaan 2345 kertaa, joten sirontasuhteeksi saadaan 64.3 %.

6 Tulosanalyysi

Luvussa tarkastellaan saatuja tuloksia, käytettyjä menetelmiä sekä työn tuloksista syntyneitä ajatuksia. Alussa pohditaan hiukan virhelähteitä ja aineiston luonnetta. Tästä jatketaan saatujen tulosten analysoinnilla. Lopuksi muodostetaan kokonaisnäkemys tutkimuksen tuottamista tuloksista.

6.1 Huomioitavat virhelähteet

Karsintaoperaation yhteydessä kiinnitettiin huomiota *entries*-tiedoston suureen redundanssiin. Samoja proteiineja ja proteiiniperheitä oli nimetty hiukan toisistaan poikkeavalla tavalla. Samoin lähdeorganismi esiintyi erilaisten määreiden jälkeen. Tämän takia aineistosta saattoi karsiutua tarkoituksetta joitakin proteiineja tai aineistoa häiritseviä proteiineja jäi mukaan.

Karsinnan yhteydessä karsiutui myös useita DNA- tai RNA-merkkijonolla leimattuja proteiineja. Sekaannusta aiheuttavat merkkijonot esiintyivät yleensä proteiinin tarkassa nimessä. Myöhemmin paljastui, että DNA- ja RNA-molekyylit voitaisiin karsia myös sekvenssin pituuden avulla. Koska proteiineja karsiutui aiheettomasti aineistosta, saattaa ennustetulokset olla heikommat.

Proteiiniperheiden välisissä identtisyysvertailuissa käytettiin identtisyysrajaa 65 %. Käytetyllä rajalla on vaikutusta myös testi- ja opetusjoukkojen väliseen sukulaisuuteen. Jos joukkojen välillä on suuri sukulaisuus, on rakenteet helpommin ennustettavissa. Rost suosittaa artikkelissaan jopa niin alhaista identtisyyslukua kuin 25 %. Tätä hän perustelee sillä, että proteiinit, joilla identtisyys on vain 5 %, saattavat olla rakenteeltaan samoja. [23]

Huomiota on kiinnitettävä myös DSSP-tiedostojen ongelmiin. Joitakin analyysissä tarvittavien proteiinien rakennetiedostoja ei ollut saatavissa ja joidenkin proteiinien rakennetiedostoissa ei ollut sekvenssiä. Aineistosta jäi puuttumaan näiden proteiinien sisältämä informaatio. Tämä vähentää aineiston monipuolisuutta.

Aminohapposekvensseissä esiintyi sellaisia merkkejä, joiden merkitystä ei tiedetty. Ongelma aiheuttaa epävarmuutta tuloksia kohtaan. Mukana oli esimerkiksi sellaisia merkkejä, joiden ei tiedetä tarkoittavan mitään tunnettua

aminohappoa (taulukossa 2 aminohapot B, X ja Z).

Sekvensseissä esiintyi myös pienillä kirjaimilla merkittyjä aminohappoja sekä huutomerkkejä. Pienet kirjaimet muutettiin isoiksi ja kaikki erikoismerkit jätettiin huomioitta. Ylimääräisten merkkien takia aineiston koodauksessa otettiin käyttöön 23 mahdollista merkkiä.

PPII-rakenteen pituuden määrittely koettiin työssä ongelmalliseksi. Mistä rakenne alkaa ja minne se loppuu, ovat kiistanalaisia kysymyksiä myös biokemistien keskuudessa. Valtaosa opetuksista tehtiin kolmen täydellisen aminohapon muodostavien rakenteiden perusteella. Käytetyllä rakenteen pituudella ei kuitenkaan näytä olevan vaikutusta ennustetarkkuuteen.

Työn yhteydessä rakennettiin useita tiedostoja lukevia-, dynaamista muistia käyttäviä- ja moduulien väliseen tiedonsiirtoon perustuvia tietokoneohjelmia. Nämä saattavat sisältää virheitä.

6.2 Aineiston luonteesta

Opetusjoukon frekvenssien avulla nähtiin, että PPII-aineistossa on muutamia kohonneita arvoja keskimmäisten aminohappojen kohdalla. Tämä paljastaa, että PPII-rakenne heijastuu jossain määrin sekvenssiin. Neuroverkon spektrin mukaan aminohappojen D, G, N P ja Y yksittäiset vaikutukset pystytään oppimaan. PPII-rakenne karttaa aminohappoja D, G, N ja Y. Vastaavasti rakenne viihtyy proliinin vaikutuksessa. Frekvenssi paljastaa myös poikkeamia ainakin aminohappojen H, L, S ja V läheisyydessä. Spektrissä nämä kaikki eivät paljastu.

Neuroverkon tiedetään näkevän ihmiselle mahdottoman monimutkaisia rakenteita. Verkot näyttävät kuitenkin törmäävän rajaan, jota enempää ei pysty oppimaan. Biokemiallisessa kirjallisuudessa kuitenkin väitetään aminohappojärjestyksen määräävän sekundaarirakenteen. Neuroverkot eivät osaa muodostaa tätä kuvausta tai jokin menetelmä on valittu väärin. Ratkaisevia valintoja voivat olla sekvenssien koodaustapa, neuroverkkoarkkitehtuuri, verkon topologia tai sekvenssin pituus.

PPII-luokan ja muita rakenteita edustavan ei-PPII-luokan välinen Hamming-etäisyys näyttää, että PPII-sekvenssit ovat omasta luokastaan lähes yhtä etäällä kuin ei-PPII-sekvenssit. Tämä tarkoittaa, että sellaisia ryppäitä, joissa on vain toisen luokan alkioita, ei hahmoavaruudessa esiinny runsaasti. Tämä näyttää olevan aineiston suurin ongelma.

Kun aineistosta poimitaan kymmenen prosenttia testimateriaaliksi, on testialkiolla kohtalaiset mahdollisuudet sijaita hahmoavaruudessa myös toisen luokan valta-alueella. Tällaiset tapaukset laskevat ennustustodennäköisyyt-

tä. Hamming-etäisyyden ja siroontumislukujen mukaan näitä on aineistossa paljon.

PPII-luokan sisällä on pieni joukko, jossa etäisyydet ovat valtaosaa pienemmät. Voitaisiinko nämä poimia suurella todennäköisyydellä ja jättää huomiotta epävarmat etäämmällä sijaitsevat tapaukset?

Aineiston luonnetta pyrittiin kuvaamaan myös informaatioteorian välinein. Menetelmä odotetusti lupaa menestyksellistä oppimista. Menestystä voidaan odottaa, koska identtisyydet ovat vastakkaisissa luokissa harvinaisia. Identtisyydet ovat harvinaisia ainakin sen takia, koska sekvenssit ovat pitkiä ja kombinatoriikan mukaan tietyn kirjainkombinaation syntymistodennäköisyys on

$$\frac{1}{20^{13}}. \tag{21}$$

Tämä on tietenkin teoreettinen alaraja todennäköisyydelle, mutta kuvaa sen äärellisen avaruuden valtavaa kokoa, jota neuroverkko jakaa osiin. Kuitenkin jokaisen mahdollisen tapauksen välinen Hamming-etäisyys vaihtelee pienellä välillä 0 - 13.

Jos ajatellaan, että tyypillisessä opetusryhmässä on 10000 tapausta. Tällöin avaruuden täyttöastetta ei voida ilmoittaa prosentteina eikä promilleina, vaan lukuna $9 * 10^{-12}$.

6.3 Tehdyt valinnat

Tutkimuksen alussa oli tehtävä valintoja, joiden luonnetta ei tarkasti osattu edeltä arvioida. Aluksi oli valittava menetelmä, jolla aminohapot muutetaan neuroverkon ymmärtämään numeeriseen muotoon. Aikaisemmissa tutkimuksissa ei oltu kyetty rakentamaan ongelmaan vaihtoehtoisia lähestymistapoja. Tässäkin työssä jouduttiin valitsemaan bittivektorikoodaus.

Työssä oli valittava myös muita muuttujia. Alussa neuroverkkoja opetettiin erilaisilla ikkunointimenetelmillä, tarkasteluikkunan pituuksilla ja piilosolmujen määrillä.

Ikkunointimenetelmä 1 havaittiin PPII-rakenteen tapauksessa parhaaksi. Menetelmää on käytetty lähes poikkeuksetta kirjallisuudessa esitetyissä sekundaariennusteissa. Tarkasteluikkunan pituudella havaittiin olevan myös merkitystä ennustustarkkuuteen. Mitä pidempi ikkuna, sen parempi ennuste. Tässä tulevat kuitenkin opetusaineiston lukumäärän asettamat rajoitteet nopeasti vastaan. Parhaaksi ikkunanpituudeksi havaittiin 13 aminohappoa.

Oppikirjat väittävät, että jokaiselle ongelmalle on olemassa optimaalisen kokoinen neuroverkko. Liian tehoton ei kykene kuvaamaan ilmiön monimutkaisuutta ja liian tehokas oppii aineistosta turhaa kohinaa. Työssä käytettiin

vain yhden piilokerroksen verkkoja. Aineiston määrä ei olisi riittänyt kahden piilokerroksen verkkoon ja niin monimutkaisia piirteitä ei uskottu aineistossa olevan. Myös perinteisissä sekundaarirakenne-ennustuksissa on käytetty yhden piilokerroksen verkkoja [23]. PPII-tapauksessa neljä piilosolmua antoi parhaan ennusteen. Tällöinkin liikuttiin jo aineiston määrän asettaman ehdon ylärajalla.

Tulokset poikkeavat hieman ryhmän Ruggiero, Sacile ja Rauch tuloksista. He eivät tutkimuksessaan huomanneet pituudella ja piilosolmujen määrällä olevan suurta vaikutusta ennustustarkkuuteen. Toisaalta heidän tutkimuksessaan käsiteltiin α- ja β-rakenteiden ennustamista satunnaisrakenteiden joukosta. [24]

6.4 Ennustustarkkuus

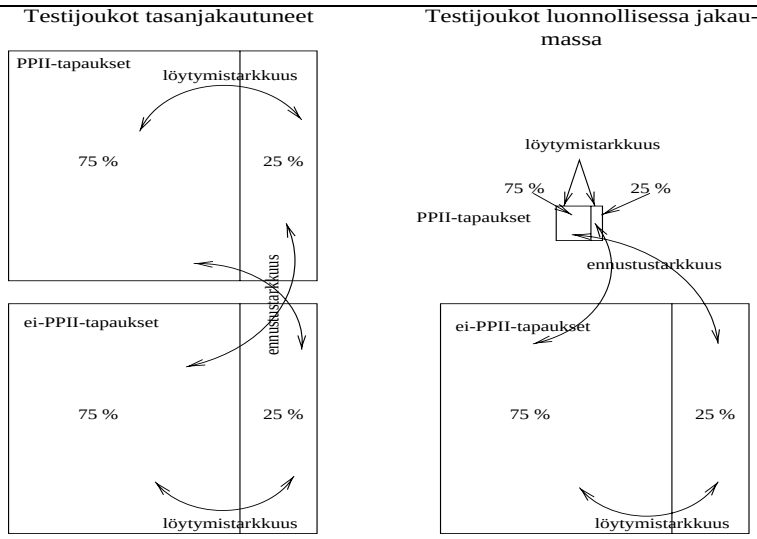
Testaus jouduttiin jakamaan kahteen eri vaiheeseen. Jotta saataisiin vertailukelpoisia ja ymmärrettäviä tuloksia, asetettiin testijoukon luokkien kokosuhteet tasapainoon. Tällöin ei-PPII-tapauksia jouduttiin karsimaan rajusti. Toisessa vaiheessa testijoukoissa käytettiin luonnossa esiintyvää jakaumaa.

Tasapainossa olevassa testijoukossa ennustustarkkuuden keskiarvo on PPII-tapauksille hieman korkeampi kuin ei-PPII-tapauksille. Kahdeksalle opetusryhmälle lasketussa keskiarvossa PPII-luokan ennustustarkkuus on 74.1 % ja ei-PPII-luokalle vastaava tulos on 73.3 %. Tuloksissa on huomioitava se, että ennustustarkkuus riippuu myös vastakkaisesta luokasta. PPII-luokan ennustustarkkuus on parempi sen takia, koska ei-PPII-luokan tapauksista löydetään suurempi osuus.

Mitkään tehostukset eivät tunnu vaikuttavan tähän tulokseen. Eräässä testissä aineistosta poistettiin vastakkaisten luokkien identtiset tapaukset. Tämä aineisto olisi saanut informaatioteoreettisesti täydellisen opittavuusarvon. Tulokset jäivät kuitenkin kahdeksan opetusryhmän keskiarvojen alapuolelle. Rakenteelle määrätty pituus ei myöskään vaikuta parantavasti ennustetulokseen. Tulokset nousevat yli 70 %:n, mutta jäävät alle yleisen (kahdeksalla opetusryhmällä saadun) keskiarvon.

Luokan ei-PPII suhteen epäonnistuneet ennusteet vaikuttavat suuresti käytännössä tapahtuvassa PPII-rakenteen ennustustehtävässä. Määritelty ennustustarkkuus ei ota huomioon, kuinka suuren osan kiinnostavista rakenteista menetelmä löytää, vaan tarkastelee löydettyjen rakenteiden suhdetta rakenteiksi luultujen lukumäärään.

Luonnollisen jakauman tapauksessa testijoukossa oli n. 29350 testialkiota. PPII-tapauksia joukossa oli 350 kpl. ja ei-PPII-tapauksia 29000 kpl. Menetelmä



Kuva 21: Tasan ja luonnollisesti jakautuneen testiaineiston löytymistarkkuuden ja ennustustarkkuuden muodostuminen. Kummastakin luokasta oletetaan, että menetelmä löytää 75 % aineistosta. Vasemmalla puolella testijoukot ovat tasapainossa ja ennustustarkkuus on löytymistarkkuuden kanssa lähes samankokoinen. Oikealla puolella on kyseessä luonnollinen jakauma. Oikeinennustettujen PPII-tapausten määrä on pahasti epätasapainossa verrattuna väärinennustettuihin ei-PPII-luokan tapauksiin.

jakaa tämän joukon samalla tavalla kuin tasanjakautuneen aineiston. Neuroverkko löytää PPII-rakenteita n. 73 % ja ei-PPII-rakenteita n. 75 %. Kun rakenteiden kokonaismäärästä eritellään nämä löydetyt rakenteet, epäonnistuu verkko 98 kertaa PPII-tapauksen kohdalla ja 7366 kertaa ei-PPII-tapauksen kohdalla. Millä todennäköisyydellä verkko ennustaa PPII-rakenteita tai montako oikeaa PPII-tapausta on PPII-tapauksiksi ennustettujen joukossa? Vastaus on 3 %. Vastaavasti voidaan kysyä, montako oikeaa ennustetta tehdään ei-PPII-luokalle. Vastaus on 99 %. Kokonaisennustustarkkuudeksi saadaan n. 73 %.

Tulokset on helppo ymmärtää katsomalla kuvaa 21. Kun testijoukot ovat yhtä suuret, on oikeinluokiteltujen PPII-rakenteiden lukumäärä hyvin tasapainossa väärinluokiteltuihin ei-PPII-tapauksiin nähden. Kun aineistot asetetaan luonnolliseen suhteeseen, on oikeinluokitelluista PPII-tapauksista muodostunut mitättömän pieni joukko suhteessa väärinluokiteltuihin ei-PPII-tapauksiin.

Laskettaessa tasanjakautuneelle aineistolle geometrinen keskiarvo saadaan

$$\sqrt{0.0331 \times 0.9955} = 0.182. \tag{22}$$

Tulos tasoittaa epätasapainossa olevia ennustuslukuja ja kokonaisennustetulos ei ole enää imarteleva.

Etsittäessä PPII-tapauksia neuroverkko epäonnistuu melko varmasti. Ongelmaa yritettiin ratkaista kahden peräkkäisen verkon menetelmällä. Menetelmä odotetusti ennustaa paremmin, mutta yhä useampi PPII-tapauksista jää löytymättä.

Osa kummankin luokan alkioista on luultavasti siroontunut hahmoavaruuteen täysin sekaisin. Toinenkaan peräkkäinen neuroverkko ei osaa erottaa luokkia toisistaan.

Toisen verkon perään voitaisiin asettaa kolmas verkko. Tällöin ehkä saataisiin uusi luokitus, joka ennustaa edellistä paremmin PPII-luokan tapauksia, mutta rakenteiden löytymisprosentti laskisi uudelleen.

Tilanne on varsin ongelmallinen PPII-rakenteen ja muiden samankaltaisten harvinaisten sekundaarirakenteiden kohdalla. Sekvenssit sisältävät vihjeitä rakenteen olemassaolosta, mutta vihje ei ole varma sääntö. Verkko oppii nämä vihjeet ja löytää sekvensseistä niitä runsaasti. Harvinaisia rakenteita on kuitenkin määritelmänsä mukaan harvassa, joten verkko tulee epäonnistumaan varsin usein. Ehkä juuri tämän takia ei harvinaisten proteiinirakenteiden ennustusyrityksiä ole tieteellisessä kirjallisuudessa esitetty.

Tapaukset ovat hahmoavaruudessa varsin sekaisin. Tämä voidaan nähdä sirontasuhteista. Neuroverkko joutuu lajittelemaan sellaista aineistoa, joissa oman luokan alkio sijaitsee lähimpänä n. 60 %:n todennäköisyydellä. Jos ajatellaan todennäköisyyttä pidemmälle, niin huomataan, että todennäköisyys kahdelle peräkkäiselle saman luokan alkiolle on laskenut jo 36 %:n. PPII-luokan ja ei-PPII-luokan välille on todella vaikeaa vetää tarkkaa rajaa.

Satunnaisesti luokittuneet tapaukset antavat siroontumissuhteeksi 0.5. Koko testijoukolle vastaava luku on 0.38. Todennäköisyys saman luokan alkion löytymiseksi on kasvanut vain hieman. Kun mitattavana on vain oikeinluokittuneet tapaukset, kasvaa rypäskoko. Edelleenkin se on varsin pieni eli vain n. 4.

Kun aineistossa on mukana vain toiseen luokkaan luokittuneet, pienenee väärinluokittuneiden ryppäiden koko lähelle lukua 1. Sirontasuhde kasvaa myös voimakkaasti. Jos joukosta valitaan satunnaisesti yksi väärinluokittunut tapaus, kuuluu tämän läheisin alkio vastakkaiseen luokkaan yli 70 %:n todennäköisyydellä.

Hahmoavaruudessa on tällaisia pieniä saarekkeita, joissa on väärän luokan alkio tai ehkä kaksi. Yleistävä neuroverkko ei ehdi tai kykene muodostamaan vaadittavaa kuvausta ennen kuin testijoukko pysäyttää opettamisen. Tässä on puolestaan syynä opetusjoukon ja testijoukon erilaisuus. Miten käy nii-

den tapausten, jotka ovat luokituksen väärällä puolella? Neuroverkko toimii periaatteidensa mukaisesti ja määrää nämä väärään luokkaan.

6.5 Syntyneet hypoteesit ja jatkotutkimusaiheet

Tulokset antavat mahdollisuuden erääseen hypoteesiin: neuroverkoilla pystytään ennustamaan PPII-rakenteita suurinpiirtein 2.62-kertaisesti satunnaiseen valintaan nähden (satunnainen valinta voidaan ajatella olevan rakenteen luonnollinen esiintymistiheys proteiinisekvenssissä).

Rostin mukaan tyypillisessä sekundaariennustuksessa on mukana 32 % α-kierrettä (DSSP:ssä H) ja 21 % β-lamellia (DSSP:ssä E) [22]. Jos näihin lukuihin sovelletaan PPII-rakenteen ennustustarkkuudesta ja esiintymistiheydestä saatua suhdetta, saadaan α-kierteelle luku 83.8 % ja β-lamellille luku 55.0 %. Voisiko näin saada harvinaisen ja yleisen sekundaarakenteen ennustetulokset vertailukelpoiseksi? Tällä tavoin ajateltuna PPII-rakennetta pystytään ennustamaan jopa paremmin kuin esimerkiksi α-kierrettä.

Suhde on myös varsin lähellä neeperinlukua e (≈ 2.718). Voisiko ennustustarkkuus olla ylhäältä rajoitettuna esiintymistiheyden ja luonnollisen logaritmin kantaluvun tulon suuruiseksi?

Tutkimuksen aikana on yhä enemmän alkanut kiinnostamaan sekundaarirakenteiden ja sekvenssien suhde yleisesti. Yhä uudelleen huomaa kysyvänsä, selittävätkö muutamien aminohappojen mittaiset sekvenssit taustalla olevat sekundaarirakenteet. Kiinnostusta lisää vielä se (tässäkin työssä kohdattu) raja, jonka yli neuroverkoilla ei näytetä pääsevän. Jos tällainen raja on todella olemassa, se lupaa huonoa tulosta PPII-rakenteille ja muille PPII-rakenteiden tapaisten harvinaisten proteiinirakenteiden ennustamiselle.

Jos neuroverkolla ei päästä tiettyä rajaa parempiin tuloksiin, mitä pitäisi tehdä? Miten voidaan ohittaa satunnaisuus, jota sekvensseissä näyttää esiintyvän. Tämä satunnaisuus sirottaa datapisteitä sekaisin hahmoavaruuteen ja vastakkaiset testijoukot sekä opetusjoukot ovat osittain päällekkäin.

Satunnaisuutta ei tunnetusti ole kaikki, mille ei löydetä matemaattista kuvausta tai mikä ei ihmisen silmissä näytä käyttäytyvän säännöllisesti. Nämä ilmiöt ovat täysin deterministisiä, mutta yhteyttä ei vain monimutkaisuudesta johtuen nähdä. Tämä ilmiö saattaa vaivata niin sekvenssejä kuin proteiinimolekyylin kolmiulotteisia rakenteitakin.

Tuloksista nähtiin, että käytetyllä tarkasteluikkunan pituudella on vaikutusta ennustustodennäköisyyteen. Pitäisikö neuroverkon käyttämät sekvenssit olla vieläkin pidempiä? Tässä vastaan tulevat neuroverkon yhteyksien määrät ja opetusaineiston puute. Pidemmät sekvenssit alkavat vaikuttamaan hidas-

tavasti myös laskentaresurssien puolella.

Jokainen sekvenssiä kuvaava bittivektori on yhtä etäällä hahmoavaruuden origosta. Aiheuttaako tämä neuroverkolle vaikeuksia, kun tämä jakaa hahmoavaruutta eri luokille kuuluviin alueisiin? Jos näin on, miten sekvenssit pitäisi koodata? Tätä kysymystä mietittiin myös tämän työn alkupuolella, eikä siihen saatu parempaa vastausta. Kirjallisuudessa ei esiinny muitakaan lähestymistapoja. Ainut poikkeus on luvussa 2 esitetty fraktaalikoodaus.

Voitaisiinko sekundaariennustuksessa käyttää koodausta, joka käyttää aminohappojen sukulaisuuksia hyväkseen. Toinen vaihtoehto voisi olla järjestää opetusvektorit siten, että niissä paljastuvat aminohappojen sivuketjujen mahdolliset sidoskumppanit.

Koodausta mietittäessä pitäisi käydä lisää biokemistien ja tietojenkäsittelytieteilijöiden välistä keskustelua. Näin menetelmästä saataisiin varmasti kummankin tieteenalan hyväksymä ja tuloksia voitaisiin odottaa.

Oppimiseen ja datapisteiden hajanaiseen sirotteluun voi osittain olla syynä myös tarkasteluikkunan liu'utus sekvenssin yli. Yhtä positiota ennen, kun PPII-rakenne alkaa, kuuluu tarkasteluikkunallinen ei-PPII-rakenteen luokkaan. Seuraava ikkunallinen kuuluukin jo PPII-rakenteen luokkaan. Näissä kahdessa sekvenssissä esiintyy vain yksi eri aminohappo. Sama toistuu, kun PPII-rakenne loppuu. Jos rakenteita on kaksi tuhatta, on tällaisia läheisiä sekvenssejä kummassakin luokassa jo neljä tuhatta.

Tarkasteluikkunan liu'utus ja ikkunointimenetelmä 1 ovat toisaalta varsin käyttökelpoisia. Näin pystytään tarkasti määrittelemään, onko verkon ennustus osunut tarkasteluikkunan keskipisteeseen. Jos verkko on oppinut ennustustehtävänsä, osataan määrätä rakenteen alku- ja loppukohdat aminohapon tarkkuudella.

Jälleen täytyy kysyä, onko harvinaisten rakenteiden kohdalla tarpeellista tietää tarkasti, mistä se alkaa ja minne se loppuu. Jos se alkaa kohdasta k, se loppuu suurella todennäköisyydellä ennen kohtaa $k + 4$. Jos ikkunaa ei liu'uteta, niin rakenteen alku voitaisiin kohdistaa sekvenssiä ikkunoitaessa tarkasteluikkunan tiettyyn positioon. Yhdestä rakenteesta saataisiin yksi opetus- tai testijoukon tapaus. Menetelmällä saataisiin huomattavasti vähemmän opetusaineistoa, joka tuo tietenkin omalta osaltaan lisää ongelmia. Verkko kuitenkin oppisi huomaamaan kahden rakenteen vaihdoskohdan. Rakenteen loppua ei edes yritettäisi ennustaa, jolloin kiinnostuksen kohteena olisi osumistodennäköisyys rakenteen alkuun.

Näiden ongelmien voittamiseksi täytyy tehdä lisää tutkimusta. Käyttöön voitaisiin ottaa jokin muu työväline kuin neuroverkot. Tässä vaiheessa on syytä mainita eräs assosiatiivimuistin kaltainen menetelmä, joka antaa

hyviä ennustetuloksia. Valitettavasti PPII-rakenteista löydetään vain noin 32 %. Menetelmä perustuu siihen, että testitilanteessa luokitellaan vain opetusjoukon kanssa samankaltaisia sekvenssejä ja pyritään välttämään liiallista yleistämistä.

Toinen lähestymistapa voisi olla pyrkimys täydelliseen kuvaukseen sekvenssien ja sekundaarirakenteiden välille tai aminohappojen ja taipumiskulmien välillä. Tämä näyttää kuitenkin olevan enemmän biokemistien kuin tietojenkäsittelytieteilijöiden työtä.

Lopuksi voidaan sanoa, että tutkimus näyttää karulla tavalla sen todellisuuden, jonka kanssa harvinaisten sekundaarirakenteiden ennustukset joutuvat kamppailemaan. Sekvensseissä oleva satunnaisuus (tai satunnaisuudelta näyttävä ilmiö) estää täydellisen kuvauksen. Tämän jälkeen ennustusta rajoittavat lait voidaan johtaa harvinaisten tapausten määritelmästä (kuten kuviossa 21 on näytetty).

Luonnollisen jakauman yhteydessä kokonaisennustetulos 73.3 % on samaa suuruusluokkaa kuin perinteisissä sekundaarirakenne-ennusteissakin [24]. Tältä osin voidaan olla tyytyväisiä. Tulos on myös parempi, mitä tilastollisilla menetelmillä ollaan saatu [23]. Neuroverkot näyttävät olevan siis paras saatavilla oleva menetelmä sekundaarirakenne-ennustukseen.

Valitettavasti proteiineisekvensseissä esiintyy liian paljon PPII-rakenteeseen viittaavia aminohappojärjestyksiä, joiden taustalla ei kuitenkaan esiinny itse rakennetta. Ilman lisäinformaatiota neuroverkko ei voi luokitella identtisiä sekvenssejä eri luokkiin.

7 Tulosten yhteenveto

Työssä on näytetty, kuinka neuroverkot pystyvät ennustamaan proteiineissa esiintyvää polyproliini II-sekundaarirakennetta. Aluksi testaus suoritettiin tasanjakautuneella testiaineistolla, jonka jälkeen testaus tehtiin luonnollisessa suhteessa esiintyville rakenteille.

Tutkimuksessa kohdattiin suuria ongelmia aineiston luonteen vuoksi. PPII-rakenteita esiintyy luonnossa harvassa, vaikka neuroverkko tarvitsee runsaasti opetusaineistoa. Aineistoa ei voi monistaa, sillä PPII-tapauksissa ja muita luokkia edustavissa sekvensseissä esiintyy runsaasti samankaltaisuutta. Opetus ja testijoukot saatettiin tasapainoon pienentämällä muiden rakenteiden luokkaa.

Neuroverkon käytölle aiheuttaa ongelmia myös vastakkaisten luokkien päällekkäisyys. PPII-rakenteita edustavissa sekvensseissä on runsaasti muita rakenteita edustavien sekvenssien piirteitä ja päinvastoin. Tämän takia sekvenssit ovat sekaisin hahmoavaruudessa ja neuroverkolla on vaikeuksia jakaa avaruutta eri luokille kuuluviin osiin.

Kun yhteen asetetaan PPII-rakenteen harvinainen esiintyminen ja päällekkäiset luokat, on tilanne hankala mille tahansa ennustusmenetelmälle.

Työn alkuosassa opetettiin verkkoja erilaisilla sekvenssien pituuksilla, kahdella ikkunointiratkaisulla sekä useilla neuroverkon piilosolmujen lukumäärillä. Näin etsittiin paras kombinaatio em. muuttujista. Parhaita tuloksia syntyi käyttämällä 13 aminohappoa pitkää sekvenssiä, neljää piilosolmua sekä tarkastelemalla PPII-rakennetta tarkasteluikkunan keskimmäisestä positiosta.

Parhaalla menetelmällä neuroverkot kykenevät löytämään PPII-rakenteista 72.6 % ja muista rakenteista 74.7 %. Tämä suhde säätelee ennustustarkkuutta niin tasanjakautuneessa aineistossa kuin myös luonnollisessa jakaumassa.

Tasanjakautuneelle aineistolle saatiin keskimääräiseksi ennustustarkkuudeksi 74.1 % PPII-rakenteille ja 73.7 % ei-PPII-rakenteille. Aineiston vino jakauma aiheutti sen, että PPII-rakenteiksi ennustettujen joukossa esiintyi 3.3 % ehdot täyttäviä PPII-rakenteita. Samoin vinon jakauman takia PPII-rakenteita joutui väärään luokkaan tosi harvoin. Tämän takia muita raken-

teita ennustettiin oikeaan luokkaan 99.5 %. PPII-rakenteiden ennustustarkkuutta voidaan parantaa, mutta rakenteita löytyy yhä vähemmän.

Ennustustarkkuus koko aineistolle on 73.3 %. Tämä tulos vastaa muiden tutkimusten tuloksia. Tavallisten sekundaarirakenteiden ennustuksissa neuroverkoilla päästään yli 70 %:n ennustustarkkuuteen.

Kun neuroverkolle annetaan syötteenä kokonaisen proteiinin sekvenssi, on erehtymistodennäköisyys verraten suuri. Syy ei ole neuroverkoissa, vaan aineistossa. Vinon jakauman lisäksi erehtymistodennäköisyyttä kasvattaa se, että rakenteelle asetetut kriteerit ovat tarkkoja. Aineistossa on kuitenkin runsaasti myös PPII-rakenteiden kaltaisia rakenteita, jotka eivät täytä PPII-rakenteen säännöllisyysehtoa tai ovat liian lyhyitä. Nämä puutteelliset rakenteet näyttävät heijastuvan sekvenssiin samalla tavoin kuin kriteerit täyttävät PPII-rakenteetkin. Nämä aineiston ongelmat vaikeuttavat PPII-rakenteen ennustamista.

Näyttäisi siltä, että neuroverkon ennustusta ei voida käyttää sellaisenaan PPII-rakenteen paikantamiseen. Avuksi täytyy saada menetelmä, joka karsii virheellisiä ennusteita. Tällaisia menetelmiä voivat olla useampi peräkkäinen verkko, assosiatiivimuisti, valmiit ennustepalvelimet tai jokin biokemiallinen lisäinformaatio.

Viitteet

[1] A Adzhubei, M Sternberg. Left-handed polyproline II helices commonly occur in globular proteins. *Journal of Molecular Biology*, 229:472 – 493, 1993.

[2] J Brown, E DeRouin. *Neural network training on unequally represented classes.* Martin Marietta Corporation, 1991.

[3] H Demuth, M Beale. *Neural network toolbox for use with Matlab.* The Math Works, 1992.

[4] P Faricelli, R Casadio. Htp: a neural network-based method for predicting the topology of helical transmembrane domains in proteins. *Cabios*, 12:41 – 48, 1996.

[5] G Fullen. A gentle guide to multiple alignment. *http://www.techfak.uni-bielefeld.de/bcd/Curric/MulAli/mulali.html (22.3.1999)*, 1996.

[6] J Hanke, J Raich. Kohonen map as a visualization tool for the analysis of protein secuences: multiple alignments, domain and segments of secondary structures. *Cabios*, 12:447 – 454, 1996.

[7] E Haug, O Sand, Ö Sjaastad, K Toverud. *Ihmisen fysiologia.* WSOY, 1992.

[8] S Haykin. *Neural Networks : A Comprehensive Foundation.* Prentice Hall, 1994.

[9] P Hietala. Tilastollisten menetelmien perusteet 1 ja 2. *Kurssimateriaali, Tilastotieteen laitos, Tampereen yliopisto*, 1996.

[10] A Hutchinson. *Algorithmic Learning.* Glearendon Press, 1994.

[11] M Juhola. Neurolaskenta. *Kurssimateriaali, Tietojenkäsittelyopin laitos, Tampereen yliopisto*, 1998.

[12] W Kabsch, C Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577 – 2637, 1983.

[13] W Katz, J Snell, M Mericel. Artificial neural networks. *Methods in Enzymology*, 210:610 – 632, 1992.

[14] M Kubat, R Holte, S Matwin. Machine learning for the detection of oil spills in satellie radar images. *Machine Learning*, 30:195 – 215, 1998.

[15] M Kubat, S Matwin. *Adressing the course of imbalancet training sets: One sided selection.* Proceedings of the Fourteenth International Conference on Machine Learning (179 - 186), 1997.

[16] E Mäkinen. *Algoritmien suunnittelu ja analyysi.* Tietojenkäsittelyopin laitos, Raportti C, Tampereen yliopisto, 1996.

[17] D Mount. Dayhoff scoring matrices (percent accepted mutation or pam matrix) for sequence comparisons. *http://www.blc.arizona.edu/courses/bioinformatics/dayhoff.html (2.8.1999)*, 1996.

[18] S Needelman, C Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48:443 – 453, 1970.

[19] M Niemi, K Korhonen, I Virtanen. *Solu- ja molekyylibiologia.* Weilin+Göös, toinen laitos, 1989.

[20] S Petersen, H Bohr, J Bohr, S Brunak, R Cotteril, H Fredholm, B Lautrup. Training neural networks to analyse biological sequences. *Tibtech*, 8:304 – 308, 1990.

[21] R Rojas. *Neural Networks : A Systematic Introduction.* Springer, 1996.

[22] B Rost. *A neural network for prediction of protein secondary structure in E Fiesler, R Beale (ed.) Handbook of Neural Computing, G4.1:1 - 12.* Oxford university press, 1997.

[23] B Rost, C Sander. 3rd generation prediction of secondary structure. *Predicting protein structure, Humana Press*, 1998.

[24] C Ruggiero, R Sacile, G Rauch. Peptides secondary structure prediction with neural networks: A criterion for building appropriate learning sets. *Transaction on Biomedical Engineering*, 40:1114 – 1121, 1993.

[25] I Suominen, P Ollikka. *Yhdistelmä-DNA-tekniikan perusteet.* Opetushallitus, 1997.

[26] K Swingler. *Applying Neural Networks a Practical Guide.* Academic Press, 1996.

[27] Z Szapo. Polyproline helices. *http://iona.cryst.bbk.ac.uk/assignments/projects/szabo/index.htm (3.8.1999)*, 1997.

[28] L Turpeenoja. *Biokemiaa, Virtsa-aineesta lääkemaitoon.* Opetushallitus, toinen laitos, 1997.

[29] I Ulmanen, J Tenhunen, J Ylänne, J Valste, P Viitanen. *Geeni.* Söderström, 1998.

[30] M Vihinen. Keskustelu polyproliinin ominaisuuksista. *Suullinen tiedonanto*, 1998.

A Identtisyysvertailun kompleksisuustarkastelu

Tarkastellaan luvussa 4.2.3 esitetyn algoritmin aikavaatimusta sekvenssien pituuksiin nähden. Kompleksisuus on algoritmin tehokkuuden mitta, joka on käytettävän tietokoneen nopeudesta riippumaton. Algoritmin kompleksisuuden tarkka laskeminen on monesti vaikeaa ja tämän takia käytetään suuntaa-antavia asymptoottisia merkintätapoja. Asymptoottisissa merkintätavoissa merkintä $f = O(h)$ tarkoittaa, että f kuuluu funktion h määräämään kompleksisuusluokkaan. Tarkasti määritellen $f = O(h)$, jos on olemassa sellaiset positiiviset vakiot c ja n_0, että

$$|f(n)| \leq c|g(n)|,$$

kun $n \geq n_0$. [16] Tässä tarkastellaan Needelmanin ja Wunchin alkuperäistä algoritmia, sillä vertailutaulukon ja kustannusfunktion mukaantuonti lisää laskutoimituksia vain kerrannaisen verran.

Kohdassa 1 käydään jokainen solu kerran läpi, jolloin tästä saadaan laskutoimituksia $m \times n$. Oletetaan, että $m < n$, niin tällöin ensimmäisen kohdan kompleksisuus on $O(n^2)$.

Vaiheessa 2 tehdään algoritmin suurin työ. Aluksi voidaan tarkastella matriisissa liikkumista. Laskenta lähtee solusta $M(2,2)$ josta edetään rivi 2 loppuun ja siirrytään riville kolme ja soluun $M(3,2)$. Tätä jatketaan, kunnes saavutaan soluun $M(n,m)$. Siis jokaisella rivillä tehdään $m-1$ operaatiota ja nämä operaatiot tehdään $n-1$ rivillä. Tehtäessä laskutoimitusta $M(i,j) = M(i,j) + max(M(i-1,..j-1), M(..i-1,j-1))$ joudutaan tarkastelemaan edellisen rivin pienempiä sarakkeita kuin j ja edellisen sarakkeen pienempiä rivejä kuin i. Summataan aluksi kaikki operaatiot, joita syntyy rivien suuntaisesti. Kuljettavia rivejä on siis $n-1$ kpl. ja yhdellä rivillä on sarakkeita $m-1$. Yhdellä rivillä tehdään

$$1 + 2 + 3 + ... + (m-1) = \frac{(m-1)m}{2}$$

operaatiota ja kaikilla riveillä tehdään yhteensä

$$(n-1)\frac{(m-1)m}{2}$$

operaatiota. Tarkastellaan sarakkeittain vastaavaa toimitusta. Kertoimeksi tulee sarakkeiden lukumäärä $n-1$ ja yhdellä sarakkeella tehdään

$$1 + 2 + 3 + ... + (n-1) = \frac{(n-1)n}{2}$$

operaatiota. Kaikkien sarakeoperaatioiden kohdalla päädytään lukumäärään

$$(m-1)\frac{(n-1)n}{2}.$$

Kun otetaan sekä rivien että sarakkeiden yksittäiset tarkastelut huomioon, saadaan operaatioiden määräksi yhteensä

$$(n-1)\frac{(m-1)m}{2} + (m-1)\frac{(n-1)n}{2}.$$

Ottamalla $n-1$, $m-1$ ja $\frac{1}{2}$ yhteisiksi tekijöiksi saadaan

$$\frac{(n-1)(m-1)}{2}(m+n)$$

ja kun oletetaan, että $m \leq n$, niin saadaan

$$\frac{(m-1)(n-1)}{2}(m+n) \leq (n-1)^2 n$$

Valitsemalla $c = 1$ ja $n_0 = 1$, niin

$$cn^3 \geq (n-1)^2 n.$$

Ylärajan määritelmän mukaan voidaan tehdä johtopäätös, että $(n-1)^2 n = O(n^3)$. Vastaavasti voidaan nähdä, että algoritmilla on asymptoottinen alaraja. Kun oletetaan, että $m \leq n$, niin alarajan kompleksisuus on $\Omega(m^3)$.

GCG-ohjelmiston toteutuksessa ollaan päädytty aproksimoivaan Monte Carlo-algoritmiin. Tämä algoritmi löytää suurella todennäköisyydellä oikean vastauksen, mutta täyttä varmuutta ei pysty takaamaan. Mitä suurempi todennäköisyys halutaan, sitä enemmän laskentaa joutuu tekemään [16].

B Neuroverkko ja opittu ilmiö

Neuroverkkoja pidetään mustina laatikkoina, jotka oppivat jollain tarkkuudella kuvauksen lähtöjoukosta maalijoukkoon. Kuvaus on tallettunut kerrosten välisiin yhteyksiin ja sitä on vaikea ymmärtää tarkastelemalla neuroverkon rakennetta tai neuroneiden välisten yhteyksien painokertoimia. Tässä aliluvussa esitetään menetelmä, jolla voidaan saada tietoa opetusdatan synnyttäneestä ilmiöstä. Menetelmässä kysytään verkolta, mitkä ovat ne muuttujat, jotka vaikuttavat voimakkaasti ilmiön syntymiseen.

Neuroverkko kykenee tunnistamaan monimutkaisuudeltaan eritasoisia ilmiöitä. Yksinkertaisimmillaan tietyn muuttujan aktiivisuus kasvattaa ilmiön todennäköisyyttä. Merkitään tällaista ilmiötä vaikeusasteella A. Kun syötteen muuttujien välillä on monimutkaisia ehtoja ilmiön olemassaololle, on kyseessä monimutkainen ongelma (vrt. XOR-ongelma). Merkitään tällaisia ilmiöitä vaikeusasteella B. Voidaan myös sanoa, että muuttujien välillä on suora tai ehdollinen riippuvuus ilmiön syntyyn.

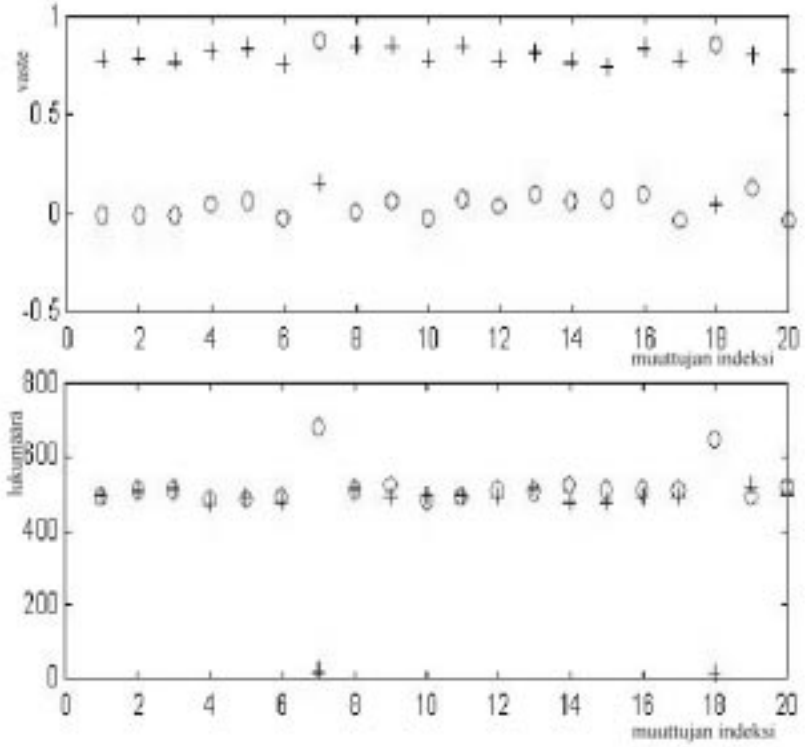
Neuroverkon "spektri"

Tarkastellaan yhden piilokerroksen perceptron-verkkoa ja oletetaan, että neuroverkon syötevektori muodostuu luvuista 0 ja 1. Tällöin oletetaan, että kukin syötevektori edustaa jotakin ilmiötä. Vektorissa luku 1 merkitsee muuttujan aktiivisuutta ja luku 0, että muuttuja ei ole aktiivinen. Menetelmä soveltuu luokitteluun ja ennustustehtäviin opetetun verkon analysointiin.

Määritellään, että neuroverkon *"spektri"* on verkon antama tulosjoukko. Tulosjoukon koko on $r \times t$, missä r on tulossolmujen lukumäärä ja t on syötteessä olevien muuttujien lukumäärä. Spektri saadaan käyttämällä syötevektorin jokaisessa muuttujassa vuorollaan lukua 1, kun muissa muuttujissa on luvut 0. Jokaisen vektorin tapauksessa mitataan tulossolmujen antama tuloste ja tallennetaan spektriin. Spektristä voidaan nähdä syötteen kunkin muuttujan merkitsevyys kullekin luokalle. Spektrillä voidaan havaita vaikeusasteen A ilmiöitä.

Spektristä saadaan paras kuva tarkastelemalla graafista esitystä (katso kuvan 22 ylempi esitys). Kuvan vaaka-akselilla on 20 ($= t$) positiota ja pystyakselilla kahden ($= r$) tulossolmun antamat vasteet. Ts. muuttujia on 20 ja tulosluokkia on kaksi. Neuroverkon toinen solmu ilmoittaa vasteen luokalle 1 (0-merkit) ja toinen solmu vasteen luokalle 2 (+-merkit). Kuvan neuroverkkoa on opetettu n. 400 opetuskierrosta.

Kuvan 22 esimerkkitapauksessa neuroverkolle on generoitu yksinkertaisia ehtoja sisältävä aineisto. Aineistoon otettiin kaksituhatta alkiota ja luokille asetetut ehdot ovat seuraavat:



Kuva 22: Neuroverkon antama spektri ja opetusaineiston muuttujakohtainen frekvenssi. Ylemmässä kuvassa neuroverkko on oppinut reagoimaan voimakkaasti muuttujien 7 ja 18 kohdalla. Samat muuttujat erottuvat myös alemmassa frekvenssijakaumassa.

1: LUOKKA 1, jos muuttuja 18 on aktiivinen tai
2: LUOKKA 1, jos muuttuja 7 on aktiivinen ja
3: LUOKKA 2 muutoin.

Kuvan alemmassa osassa on vertailun vuoksi esitetty generoidun aineiston molempien luokkien frekvenssit. Kuvia vertailemalla nähdään, että neuroverkko on oppinut reagoimaan voimakkaimmin juuri niiden muuttujien kohdalla, jossa on frekvenssin suurimmat erot, eli muuttujien 7 ja 18 tapauksiin. Näiden muuttujien kohdalla luokkaa 1 edustava solmu on antanut huomattavasti voimakkaamman vasteen kuin toista luokkaa edustava solmu. Juuri näiden muuttujien aktiivisuus vaikuttaa suuresti neuroverkon päätökseen. Näin kuuluu tehdäkin, sillä tämä on ehtona generointisäännöissä. Vastaava ilmiö esiintyy voimakkaasti myös alemmassa frekvenssikuvassa; neuroverkko siis löytää samat ilmiöt kuin aineistolle suoritettu frekvenssianalyysi. Muuttujien yhteisvaikutus ei kuitenkaan paljastu.

Neuroverkon vasteanalyysi

Tarkastellaan usean muuttujan aiheuttamia monimutkaisia vuorovaikutussuhteita. Näitä vaikeusasteen B ilmiöitä ei voi havaita tarkastelemalla pelkästään opetusdatan frekvenssejä.

Tämän voi ymmärtää ajattelemalla 2-ulotteista XOR-ongelmaa. Tällöin luokan P syötteet ovat vektoreita $(0,0)$ ja $(1,1)$ sekä luokan N syötteet ovat vektoreita $(0,1)$ ja $(1,0)$. Laskettaessa luokkien muuttujien frekvenssit saadaan kummastakin luokasta samat jakaumat. Ilmiötä ei välttämättä voi havaita tarkastelemalla aineistoa. Monimutkaisia suhteita ei ymmärretä tai aineistossa on liikaa tietoa tai häiritsevää kohinaa.

Vasteanalyysillä pyritään paljastamaan kaikki se tieto, mitä neuroverkko on opetusaineistosta oppinut.

Algoritmi olettaa, että on olemassa eteenpäinsyöttävä neuroverkko, joka on oppinut jostain ilmiöstä olennaisia piirteitä ja pystyy luokittelemaan ilmiön syötetapauksia. Toimintaa voidaan kuvata seuraavan algoritmin mukaisesti:

1: GENEROINTI: Generoidaan neuroverkolle sallittuja syötekombinaatioita ja valitaan syötekombinaatioista tarkasteltavalle ilmiölle korkeita vasteita antavia tapauksia (muuttujakombinaatioita).
2: SAMMUTUS: Yksittäisen muuttujakombinaation jokainen aktiivinen muuttuja käydään läpi. Aktiivinen muuttuja asetetaan passiiviseksi ja tarkastetaan, laskeeko neuroverkon vaste oleellisesti. Jos laskee, muuttujan aktiivisuus on tärkeä tässä kombinaatiossa. Jos vaste ei laske, ei muuttujalla ole merkitystä ilmiön syntyyn tässä kombinaatiossa ja muuttuja voidaan jättää passiiviseksi (sammuttaa).
3: KLUSTEROINTI: Etsitään muuttuja-avaruudesta paikallisia ryhmittymiä ja muodostetaan ryhmittymästä yksi ryhmää kuvaava alkio (keskialkio).
4: TULKINTA: Keskialkioista yritetään muodostaa kuva ilmiön synnylle tärkeistä muuttujakombinaatioista.

Generointivaiheessa haetaan suhteellisen suuri joukko syötteitä, jotka verkko tunnistaa ilmiön aiheuttajiksi. Tällöin on hyvä, jos generointi voi pysähtyä myös paikalliseen optimiin. Jos optimeja löytyy runsaasti, on ilmiöllä runsaasti erilaisia synnyttäviä kombinaatioita. Liian tehokas etsintä johtaa vain parhaiden alkioiden havaitsemiseen ja ilmiön monimuotoisuus voi jäädä huomaamatta. Generointiin voidaan käyttää esimerkiksi geneettisiä algoritmeja.

Syötekombinaatiot sisältävät muuttujia, jotka jäävät aktiivisiksi sattumalta. Voi olla, että niillä ei ole merkitystä ilmiön syntyyn. Nämä muuttujat sammutetaan käyttäen uudelleen hyväksi neuroverkon antamaa vastetta. Syöt-

teestä pyritään jättämään aktiivisiksi ainostaan ne muuttujat, jotka verkon mielestä vaikuttavat ratkaisevasti ilmiön syntyyn.

Jos ilmiö on monimutkainen, hyviä vasteita antavien syötteiden joukko saattaa olla iso. Tästä joukosta pitäisi vielä saada esiin ilmiön piirteet. Joukossa voi olla kombinaatioita, jotka ovat identtisiä, samankaltaisia ja totaalisesti erilaisia. Identtiset ja läheiset alkiot voidaan ryvästää ja laskea näille jonkinlainen keskimääräinen tapaus. Muuttuja-avaruuteen voidaan synnyttää ryppäitä esimerkiksi jollakin etäisyysalgoritmilla. Lopullinen ryppään keskialkio on verkon tulkinta ilmiön synnyttämästä syötekombinaatiosta.

Menetelmän käyttäjä saa nähtäväkseen joukon keskialkioita, joista pitäisi pystyä muodostamaan tulkinta ilmiöstä. Tähän ei ole enää algoritmista lähestymistapaa, vaan kaikki on kiinni tulkitsijan kekseliäisyydestä ja aikaisemmasta tiedosta ilmiön luonteesta. Seuraavassa esimerkissä havainnollistetaan menetelmän kulkua ja mitä ilmiöstä voidaan keskialkioiden perusteella sanoa.

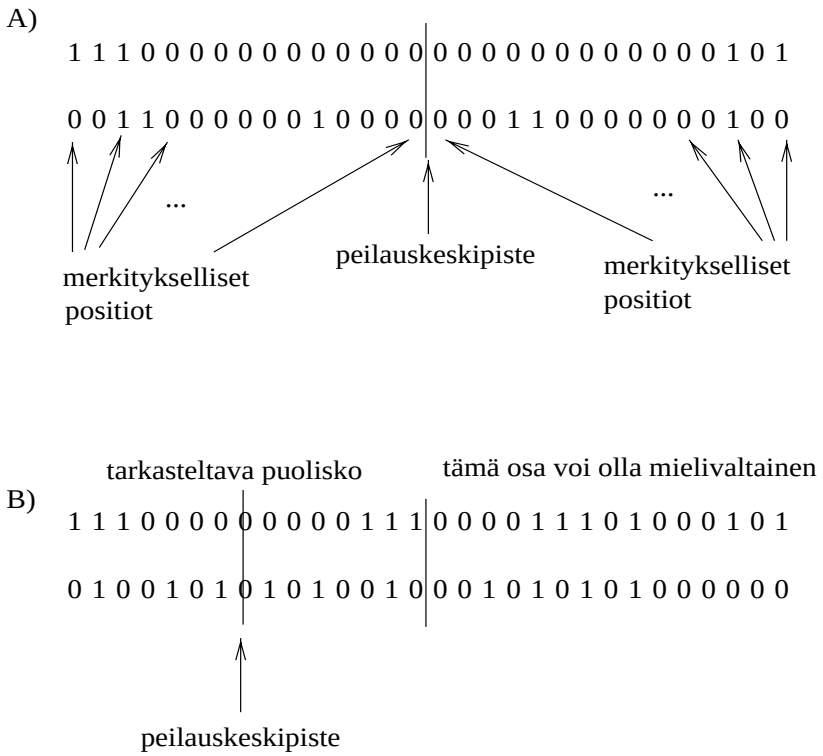
Generoidaan neuroverkolle opetusaineisto ja opetetaan yhden piilokerroksen perceptron-neuroverkko tunnistamaan ilmiö. Ongelman pitää olla sellainen, että se ei paljastu spektrissä, vaan siinä toteutuu usean muuttujan vuorovaikutussuhteet. Tehtävää monimutkaistetaan vielä siten, että generoidusta opetusalkiosta ei huomata ilmiön olemassaoloa ilman ehdon tuntemista. Tällöin menetelmän tarkoitus on tuoda esille ne piirteet, jotka oleellisesti vaikuttavat ilmiön syntyyn.

Esimerkkiongelmaksi soveltuu hyvin *peilikuvan* tunnistaminen. Neuroverkon tehtävänä on tunnistaa, onko syötetty bittivektori keskipisteen suhteen peilikuva vai ei. Peilikuvasta tehdään kaksi eri variaatiota. Toinen on sellainen, jossa peilauksen keskipiste on vektorin puolessa välissä. Peilikuvaa tarkastellaan vain joka toisen muuttujan tapauksessa. Toisessa tapauksessa keskipiste on vektorin ensimmäisen puolikkaan keskipisteessä ja tarkastellaan ensimmäisen puolikkaan jokaista alkiota. Ehdon täyttävät syötevektorit voivat saada esimerkiksi kuvan 23 kaltaisia alkioita.

Kummankin tapauksen yhteydessä generoitiin tuhat peilikuvaa ja saman verran alkioita, jotka eivät olleet peilikuvia. Yksittäisen syötevektorin kooksi valittiin 30 muuttujaa. Numero 1 kertoo muuttujan olevan aktiivinen ja 0 muuttujan olevan passiivinen. Puolipeiliongelman frekvenssi ja spektrikuvat ovat kuvassa 24. Kokopeiliongelman kuvaajat ovat lähes vastaavia.

Aineiston frekvenssistä ei voida erottaa mitään säännollisyyttä tai ilmiötä kuvaavaa ominaisuutta. Samoin spektrikin näyttää vain, että eri puoliskoilla verkko reagoi yksittäiseen aktiiviseen muuttujaan eri tavalla; spektri ei kuitenkaan kerro, miksi kuvaajan eri puoliskot ovat erilaisia.

Algoritmin toimintaa ohjaavat useat erilaiset parametrit. Geneettisen al-

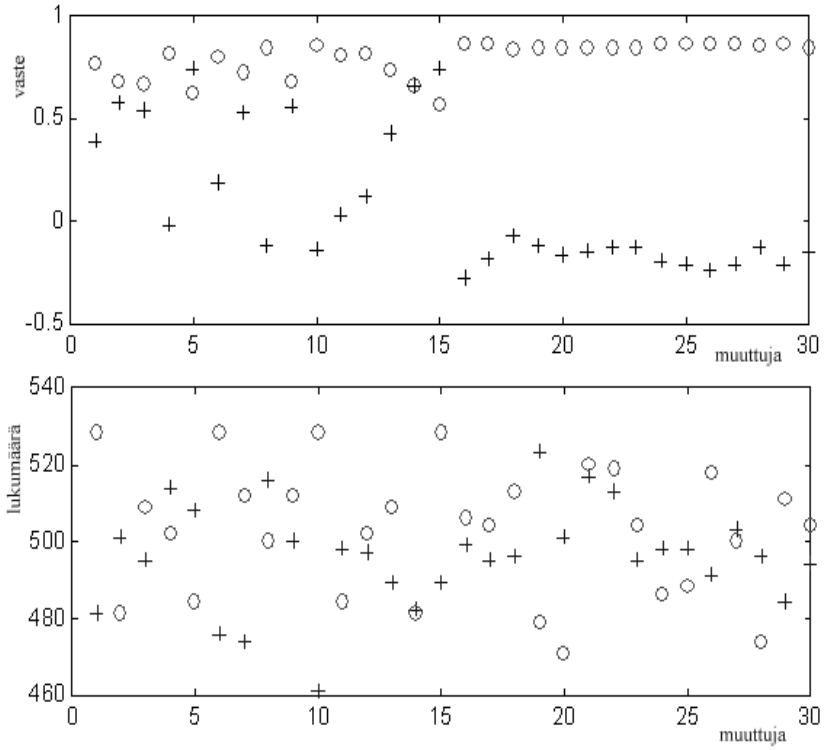


Kuva 23: Peilikuva-tehtävän esimerkkivektoreita. Tapauksen A alkiot kuvaavat mahdollisia peilikuva-alkioita kokopeiliongelmalle, jossa joka toinen alkio on merkityksellinen. Tapauksen B alkiot esittävät puolipeiliongelmaa, jossa peilauksen keskipiste on ensimmäisen puoliskon puolessa välissä.

goritmin parametreillä säädellään populaation kokoa, iteraatiokierroksia ja risteytystodennäköisyyttä. Sammuttamisvaiheessa parametrin avulla päätellään sammuttamisen tarpeellisuus. Jos muuttuja ei ole välttämätön, se sammutetaan. Ryvästysvaiheessa etäisyysparametrilla tarkastellaan muuttujakombinaation samankaltaisuutta. Ryväskoko-parametrilla säädellään pienintä sallittua ryväskokoa, josta keskialkio otetaan lopulliseen tulosjoukkoon.

Esimerkin yhteydessä populaation kokona käytettiin kymmentä alkiota, iteraatiokierrosten lukumääränä kymmenen ja risteytystodennäköisyytenä 0.2. Sammutusparametri asetettiin luvuksi 0.3 ja parhaaksi ryväskooksi osoittautui kolmen kokoinen muuttujakombinaatioiden joukko. Etäisyydeksi otettiin kolme muuttujaa.

Geneettinen algoritmi muodosti kymmenen alkion alkupopulaation ja jalosti tätä kymmenen sukupolven ajan. Jalostuksessa risteytystodennäköisyys jokaisen muuttujan kohdalla oli 0.2. Parhaalle yksilölle tehtiin tämän jälkeen muuttujien sammutus. Sammutus suoritettiin, jos sammuttaminen pienensi tulossolmujen erotusta vähempi kuin 0.3. Kun yksilöitä oli riittävästi,

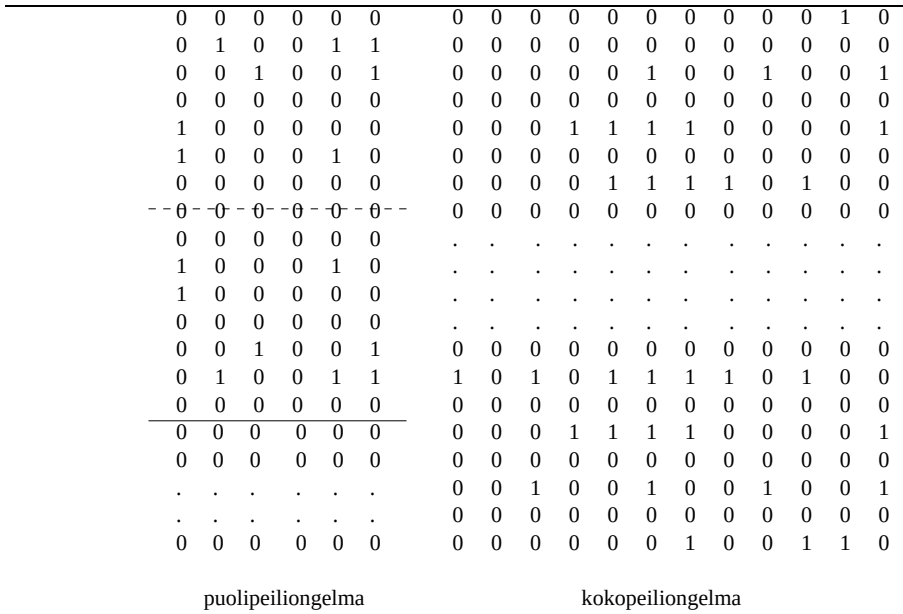


Kuva 24: Neuroverkon spektri (ylempi kuvaaja) ja opetusaineiston frekvenssi. Vaikka puolipeiliongelma kuuluukin vaikeusasteen B ongelmiin, erottuu spektristä, että verkko reagoi eri tavalla syötevektorin alkupäähän ja loppupäähän. Huomioitavaa on myös se, että frekvenssikuvaajan pystyakseli on katkaistu.

suoritettiin ryvästys etäisyyden 3 mukaan. Jos ryppääseen saatiin tarpeeksi alkioita, laskettiin ryppäälle keskialkio.

Vasteanalyysissä generoitiin 200 muuttujakombinaatiota kumpaankin peiliongelmaan, jotka kävivät läpi vasteanalyysin vaiheet. Toimenpide tuotti kuvan E keskialkiot.

Puolipeiliongelmaan vasteanalyysi tuotti kuusi täydellistä keskialkiota (kuva E). Huomioitavaa on se, että jokaisessa vektorissa viimeiset 15 muuttujaa ovat asettuneet nollaksi. Tästä voidaan päätellä, että ilmiön synnyttämät muuttujat sijaitsevat ylemmässä puoliskossa. Samoin voidaan huomata, että on myös muita muuttujia, jotka ovat jokaisessa vektorissa kokonaan passiivisia. Tästä voisi ajatella, että nämä muuttujat eivät ole ilmiölle tärkeitä. Tämä olisi väärä tulkinta. Opetusaineiston generoinnissa tai vasteanalyysin muuttujakombinaatioiden generoinnissa ei ehkä ole sattunut syntymään tarpeeksi tällaisia kombinaatioita. Keskialkioiden tarkoitus olisikin synnyttää tulkitsijalle ajatus peilauksesta, jolloin ajatusta voisi testata esi-



Kuva 25: Vasteanalyysin tuottamia keskialkioita pystyvektoreina. Vasemmalla on saatu puolipeiliongelmasta kokonaisuudessaan kuusi täydellistä tapausta, joista jokainen toteuttaa asetetun ehdon. Oikealla kokopeiliongelmasta saadut keskialkiot sisältävät enemmän kohinaa ja tänne on sattunut muutama, joka ei ole täydellinen peilikuva. Huomioitavaa on myös se, että neutraalit muuttujat on sammutettu ja alkiot kertovat selvemmin ilmiön syntyyn vaikuttavat muuttujat.

merkiksi oikean aineiston suhteen.

Kokopeiliongelmassa on vaihtoehtoja runsaasti enemmän, mutta analyysi löysi viisi alkiota, jotka ovat täydellisiä peilikuvia, viisi alkiota, jotka poikkeavat peilikuvasta yhden muuttujan verran ja kaksi alkiota, jotka poikkeavat peilikuvasta kolmen muuttujan verran. Kaikki parillisissa positioissa olleet muuttujat ovat asettuneet passiivisiksi - niin kuin pitääkin.

Tuloksista voidaan päätellä, että menetelmä korostaa tehokkaasti ne muuttujat, jotka neuroverkko on oppinut tärkeiksi. Menetelmä karsii myös menestyksekkäästi ne muuttujat, jotka eivät ole tärkeitä ilmiön synnylle. Opetusaineiston laatu on kuitenkin oleellisen tärkeää tämänkin analyysin onnistumiselle.

C Tapausten sironta

Informaatioteoreettinen tarkastelu ei kiinnitä huomiota tapausten sijoittumiseen hahmoavaruudessa. Seuraavaksi tarkastellaan työn yhteydessä syntynyttä ajatusta havaita neuroverkolle edullisia ongelmia. Menetelmä etsii etäisyyden avulla paikallisia ryhmittymiä hahmoavaruudesta. Tällöin paljastuu myös informaatioteoreettisesti hankalat aineistot.

Oletetaan, että käytössä on aineisto, jossa on kahteen eri luokkaan kuuluvia alkioita. Aluksi tapausten joukosta valitaan satunnaisesti aloitustapaus. Tämän jälkeen jonkin metriikan avulla etsitään tapauksen läheisin alkio ja aloitustapaus poistetaan joukosta. Jos lähimpiä tapauksia on useita, valitaan siirtyminen näiden joukossa satunnaisesti. Näin siirrytään aina uuteen läheisimpään pisteeseen ja poistetaan edellinen tapaus joukosta. Siirryttäessä uuteen pisteeseen lisätään laskuria a, jos uusi tapaus on eri luokasta. Muutoin a jää ennalleen. Näin käydään koko joukko läpi ja tuloksena on vaihteluiden kokonaismäärä a.

Samalla voidaan kerätä myös syntynyt polku, joka on luettelo alkioiden luokista (järjestys on tärkeä). Polusta voi myöhemmin laskea aineistoa kuvaavia tunnuslukuja.

Aineistossa, jossa luokat ovat täysin sekaisin, on paljon vaihtelua eli luku a kasvaa suureksi. Jos luokat ovat sijoittuneet avaruuteen paikallisesti eri alueille, on vaihtelua vähän. Itse asiassa vaihtelua saattaa parhaassa tapauksessa olla vain $k-1$ kertaa, missä k on luokkien lukumäärä.

Merkitään, että m on pienemmän luokan alkioiden lukumäärä ja k on luokkien lukumäärä. Tällöin pisteiden sijoittumista voidaan kuvata sirontasuhteella

$$\frac{a-(k-1)}{2m}. \tag{23}$$

Jos luokat ovat samankokoisia, tulee nimittäjään $2m-1$, jossa m on toisen luokan alkioiden lukumäärä.

Menetelmä voidaan yleistää koskemaan tapauksia, joissa luokkia on n kappaletta, jolloin suhde saadaan muotoon

$$\frac{a-(k-1)}{2m_2+\sum_{i=3}^{n} m_i}, \tag{24}$$

missä luokat i on järjestetty suurimmasta pienimpään siten, että suurin on indeksillä 1 jne. Luku m_i on luokan i alkioiden lukumäärä. Tällöin suurimman luokan lukumäärällä ei ole merkitystä vaihteluiden maksimimäärään.

Suhde saa lähelle nollaa olevia arvoja, jos pisteet ovat sijoittuneet hahmoavaruuteen edullisesti. Menetelmä antaa lähellä lukua 1 olevia arvoja,

jos alkioita valitaan vuorottain eri luokista. Menetelmän antamia arvoja tarkastellaan tuloksia käsittelevässä luvussa.

D Runsaasti PPII-rakenteita sisältäviä proteiineja

Tässä liitteessä esitellään karsinnan jälkeen aineistoon jääneitä proteiineja, joista löytyi runsaasti PPII-rakennetta. Rakenteet etsittiin Adzhubein ja Sternbergin esittämillä menetelmillä (katso lähdeluettelo). Käytettyjä PPII-rakenteen ehtoja on esitelty luvuissa 4.3 ja 4.4. Rakenteen pituutena on käytetty kolmea ehdot täyttävää aminohappoa.

Esiintymistiheys 10 % tai suurempi (9kpl.)

1a21 1a3j 1ag7 1cag 1cks 1ext 1mmc 1omc 2fdn

Esiintymistiheys 5 - 10 % (71 kpl.)

1ae5 1aht 1ai8 1an1 1aru 1azz 1beo 1bfs 1bft 1bnb 1boy 1btk 1cfb 1cpo 1cwe 1dxg 1elg 1elt 1exf 1fle 1fon 1fxy 1har 1hbt 1hic 1hne 1hoe 1hpt 1irk 1jsg 1lcj 1ldt 1llp 1mct 1mhl 1mml 1mrk 1ncg 1obs 1oya 1pce 1pgs 1pij 1pkr 1pnf 1ppc 1psp 1ptq 1sce 1slu 1stm 1tbn 1tgs 1tgx 1ton 1tpp 1try 2cga 2hft 2pf1 2pka 2psp 2tgf 2tgp 351c 3chb 3est 3pcg 3rp2 5ptp 6fd1

Esiintymistiheys 2 - 5 % (319 kpl.)

1a0n 1a1n 1a1r 1a1x 1a26 1a2v 1a2z 1a3r 1a58 1a68 1aac 1aar 1abo 1abr 1ads 1aec 1aer 1aew 1agj 1agq 1ah1 1ahc 1ahq 1aij 1aim 1ak4 1ake 1akl 1amm 1anu 1aoe 1aoh 1aoj 1aol 1aop 1aoz 1ap6 1apa 1aq6 1aqd 1aqt 1ars 1ash 1ata 1aui 1av4 1avy 1ax01ax4 1aya 1aym 1bdb 1bec 1beg 1bet 1bfd 1bfg 1bgp 1bif 1bk4 1bp2 1bra 1btg 1btn1byb 1ca0 1cew 1cfr 1cgh 1chg 1cjl 1cle 1cnt 1coy 1csb 1csn 1cum 1dan 1dbr 1dco1dhp 1dlc 1dun 1eaf 1eap 1ecf 1ecl 1ecy 1efv 1epm 1esc 1fba 1fil 1fmb 1fna 1foy1fro 1frp 1fus 1fvk 1gar 1gc1 1gct 1ghu 1gof 1gsh 1guq 1han 1hcg 1hcl 1hfi 1hfs1hia 1hja 1hnf 1hpi 1hrj 1hrn 1hum 1hus 1hvq1hxn 1hxp 1ids 1iea 1ifs 1ift 1iib 1ikf 1ilr 1ir3 1isa 1itb 1ixh 1jac 1jdw 1jud 1jvr 1jxp 1kba 1kbc 1ksi 1kst 1kvo1lam 1lbe 1lis 1lkk 1lli 1lpp 1lt5 1lxd 1mba 1mka 1mla 1mlc 1mn1 1mpp 1msp 1mvp 1nal 1nba 1nfp 1nir 1nox 1np4 1nsc 1ntn 1nxb 1oac 1obw 1omd 1onc 1onr 1opr 1opy 1ose 1otg 1oxo 1p01 1p12 1pcn 1pda 1pig 1plc 1pne 1poa 1ppe 1ppf 1ppg 1prc 1prx 1pty 1pvd 1pya 1rcn 1rdl 1rds 1rfa 1rfs 1rhs 1rom 1rot 1rp1 1rro 1rss 1rsy 1rth1rtp 1rtu 1rva 1sac 1sdf 1sfp 1sha 1skz 1slt 1smd 1smt 1snp 1sox 1ste 1stf 1tal 1tca 1tcs 1tet 1tgn 1the 1tht 1tii 1tmy 1toh 1tpk 1tpl 1trg 1tuc 1tul 1tvd 1tys 1uae 1v39 1wab 1wba 1vde 1vew 1vfn 1wgj 1vhh 1who 1wht 1vig 1vls 1vmo 1vnc 1vvc 1wyk 1xjo 1yal 1yer 1yfo 1ypt 2aak 2ait 2ak3

2alp 2asi 2bb2 2ccy 2cdv 2cel 2chs 2cnd 2csn 2cy3 2cyp 2def 2dtr 2ebn 2erk 2fbj 2gar 2gmf 2h1p 2hrp 2jel 2jxr 2lhb 2mhr 2nac 2pii 2ptl 2rhe 2sak 2sfa 2sn3 2srt 3aky 3bcl 3chy 3ebx 3lck 3pte 3rub 3seb 3ssi 3vub 4ake 4ape 4cpv 4est 4gcr 4hck 4pgm 5pal 6cel 7aat 8dfr 8fab

E PPII-rakenteissa esiintyvien aminohappojen lukumäärien suhde vastakkaisen luokan aminohappojen lukumääriin

Aineiston ominaisuuksin käsittelyn yhteydewssä esiteltiin PPII-rakenteiden ja ei-PPII-rakenteiden absoluuttiset aminohappojen lukumäärät eräällä opetusaineistolla. Tämän liitteen taulukossa on esitelty vastaavat luvut suhteellisina lukuina, kun PPII-luokan sekvensseissä esiintyviä aminohappomääriä verrataan ei-PPII-luokan vastaaviin lukuihin. Tarkasteluikkunan pituus on ollut 13.

Taulukko 16. PPII-luokan sekvensseissä esiintyvien aminohappojen lukumäärän suhde ei-PPII-luokan vastaaviin lukuihin.

A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	X	Y	Z
0.83	12.00	0.57	0.82	0.87	0.86	1.01	0.92	0.89	1.11	0.97	0.77	1.17	1.41	1.03	1.14	1.24	1.01	0.97	1.08	0.44	0.95	-
0.78	1.18	0.49	0.99	0.79	0.80	1.19	0.94	0.98	1.00	0.92	0.86	1.05	1.45	0.95	1.29	1.24	0.94	0.86	1.23	0.66	1.02	2.00
0.80	1.83	1.29	0.78	0.80	0.88	1.45	0.88	0.93	0.91	0.83	0.74	0.83	1.27	1.02	1.37	1.17	1.05	1.00	0.90	0.66	0.94	-
0.77	1.00	0.68	0.68	0.97	1.33	1.19	1.01	0.93	0.96	0.94	0.85	0.81	1.25	1.09	1.23	0.95	1.11	1.10	0.75	1.40	1.01	-
0.82	1.83	0.87	0.55	0.89	1.29	0.92	0.83	0.94	1.05	0.97	0.91	0.54	1.92	1.07	1.24	0.87	1.26	1.17	0.77	1.40	1.11	-
1.02	2.00	0.83	0.45	0.84	1.06	0.34	0.80	1.04	0.82	1.34	0.68	0.51	3.07	1.21	1.11	0.81	1.21	1.38	0.66	0.58	0.80	-
0.89	1.71	0.60	0.74	0.84	0.77	0.28	0.79	0.86	0.92	1.15	0.60	0.60	4.40	0.97	1.15	0.99	1.16	1.11	0.71	0.83	0.51	-
1.11	1.42	0.43	0.92	1.03	0.60	0.34	0.55	0.70	0.96	1.11	0.58	0.70	4.20	1.03	0.85	1.22	0.89	0.87	0.64	0.45	0.39	-
0.98	0.42	0.44	1.21	1.06	0.59	0.65	0.64	0.59	0.81	0.83	0.57	1.09	3.08	0.95	1.08	1.63	0.86	0.80	0.56	0.55	0.63	-
0.86	0.40	0.60	1.23	1.33	0.57	0.96	0.86	0.50	0.91	0.83	0.67	1.09	2.27	1.36	1.02	1.40	1.03	0.67	0.64	0.75	0.69	-
0.93	1.50	0.86	1.24	1.12	0.89	1.07	0.83	0.74	0.85	0.78	0.83	1.21	1.39	1.12	0.88	1.28	1.04	0.74	1.24	0.50	0.88	-
0.86	5.66	0.65	1.15	1.32	0.73	0.99	1.10	0.70	0.88	0.90	0.81	0.97	1.49	1.22	0.98	1.19	1.04	0.90	0.87	0.71	0.95	-
0.95	2.12	0.90	1.08	1.19	0.85	0.92	1.07	0.84	0.90	0.88	0.87	0.95	1.26	1.33	0.93	1.20	0.97	0.87	1.48	0.66	0.96	-