कृषि जैव सूचना में टूल्स और तकनीकियों का अवलोकन

ऑन-लाईन हिंदी कार्यशाला

ई - संदर्भ संहिता

पाठ्यक्रम समन्वयक : डॉ. सुधीर श्रीवास्तव

पाठ्यक्रम सह-समन्वयक : डॉ. मो. समीर फारूकी

डॉ. कृष्ण कुमार चतुर्वेदी

दिसम्बर 14-16, 2020

कृषि जैवसूचना केन्द्र (केबिन)

भा.कृ.अनु.प.–भारतीय कृषि सांख्यिकी अनुसंधान संस्थान (भा.कृ.अनु.प.–भा.कृ.सां.अनु.सं.)

लाइब्रेरी अवेन्यू, पूसा, नई दिल्ली–110012

http://cabgrid.res.in/cabin; http://iasri.icar.gov.in

आमुख

जैव सूचना विज्ञान वस्तुतः जीव विज्ञान, कंप्यूटर विज्ञान और सांख्यिकी का अंतःविषय क्षेत्र है। पिछले दो दशकों के दौरान जैविक विज्ञान के क्षेत्र में बृहद डेटा (आंकड़ा) उत्पन्न किया गया जिसमें सबसे पहले जीवों के जिनोम अनुक्रमण की विषय में जानकारी प्राप्त की गई। इसके उपरान्त इन प्राप्त जानकारियों को उच्च प्रयोगात्मक तकनीक से जैव प्रौद्योगिकी अनुसंधान प्रयोगशालाओं में किये गये प्रयोगों तथा इसके प्रभावों की गतिशीलता का अध्ययन किया जा रहा है। जैविक अनुसंधान के क्षेत्र में विभिन्न जैवसूचना विज्ञान तकनीकों/ टूल्स के प्रयोग, डेटा की संचयन एवं पुनःप्राप्ति, विश्लेषण, एनोटेसन और परिणाम के अपनी सम्पूर्णता में जैविक प्रणालियों को बेहतर ढंग में समझने में सहायक है। इन तकनीकियों के द्वारा आनेवाले समय में कृषि जैव सूचना हेतु आवश्यक सामग्री, उपकरण एवं तकनीकों के विकास को बढ़ाने में सहायक हो रहा है। इस कार्यशाला का उद्देश्य जैव सूचना विज्ञान का एक अवलोकन तथा परिषद के संस्थान एवं कृषि विश्वविद्यालय के अंतर्गत फैकल्टी, वैज्ञानिक तथा तकनीकी अधिकारियों को जैव सूचना तकनीक से अवगत कराना है। यह कार्यशाला मुख्य रूप से जीनोमिक्स, ट्रान्सक्रिप्टोमिक्स, मेटाजीनोमिक्स और प्रोटिओमिक्स के मॉडयूल पर आधारित है। इस कार्यशाला कों निम्न उद्देश्यों को ध्यान में रखकर बनाया गया है:

- प्रतिभागियों को कृषि में जैव सूचना विज्ञान के अनुप्रयोग से अवगत कराना।
- कृषि में जैव सूचना विज्ञान के विभिन्न टूल्स और तकनीकियों का अवलोकन कराना।

इस कार्यशाला में (1) अशोका का परिचय, (2) जैव सूचना विज्ञान का परिचय, (3) NGS आंकड़ों की असेंबली एवं एनोटेशन, (4) आर.एन.ए. सेक आंकड़ों का विश्लेषण, (5) एस.एन.पी. और एस.टी.आर. मार्कर विश्लेषण, (6) प्रोटीओमिक्स एवं मेटाजिनोमिक्स आंकड़ों का विश्लेषण, तथा (7) प्रोटीन संरचना मॉडलिंग को शामिल किया गया है।

यहाँ हम निदेशक महोदय, भा.कृ.अ.प.-भा.कृ.सां.अ.सं., प्रधान प्रभाग (कृषि जैव सूचना केंद्र) तथा हिन्दी एकक के प्रति आभार व्यक्त करते हैं जिनके सहयोग से इस कार्यशाला का आयोजन किया जा रहा है।

लेखकगण

अनुक्रमणिका

क्रमांक	विषय	वक्ता	पृष्ठ
1	हिंदी फोंट और हिंदी यूनिकोड का परिचय	श्री उमेश चंद्र बंदूनी	1 – 2
2	कृषि में जैविक डेटा विश्लेषण के लिये जैव कंप्यूटिंग पोर्टल, टूल्स और एल्गोरिदम	डॉ. अनिल राय	3 – 12
3	अशोका - एक परिचय	डॉ. कृष्ण कुमार चतुर्वेदी	13 – 18
4	बेसिक लोकल एलाइनमेंट सर्च टूल	डॉ. शशि भूषण लाल	19 – 24
5	जीनोम असेंबली	डॉ. द्विजेश चंद्र मिश्र	25 – 33
6	जीनोम एनोटेशन	डॉ. संजीव कुमार	34 – 47
7	जीनोमिक चयन के लिए सांख्यिकीय तरीके	डॉ. नीरज बुढलाकोटी	48 – 53
8	डी.एन.ए. सिग्नेचर आधारित एस.एन.पी. और एस.टी.आर. मार्कर विश्लेषण	डॉ. मीर आसिफ इकबाल	54 – 66
9	आर.एन.ए. सेक डेटा विश्लेषण	डॉ. मो. समीर फ़ारूक़ी	67 – 70
10	कृषि में मेटाजीनोमिक्स की भूमिका	डॉ. अनु शर्मा	71 – 77
11	कृषि में प्रोटिओमिक्स डेटा विश्लेषण का अवलोकन	डॉ. सुधीर श्रीवास्तव	78 - 84
12	प्रोटीन संरचना मॉडलिंग	डॉ. यू. बी. अंगड़ि	85 - 93

हिंदी फोंट और हिंदी यूनिकोड का परिचय

प्राप्त ई-मेल/ई-मेल के माध्यम से प्राप्त पत्र का ई-मेल से उत्तर देने के लिए -

महोदय/महोदया/ प्रिय ..... आदरणीय/,

- उपरोक्त विषय के संबंध में अपेक्षित जानकारी पत्र के साथ संलग्न/मेल-सूचना इस ई/है।
- उपरोक्त विषय के संबंध में सूचना संलग्न है।
- उपरोक्त विषय के संबंध में कृपया अनुलग्नक का अवलोकन करें।
- उपरोक्त विषय के संबंध में इस संस्थान से संबन्धित सूचना संलग्न है।
- उपरोक्त विषय के संबंध में अपेक्षित आंकड़े संलग्न हैं।

सधन्यवाद ।) साभार । /आवश्यकतानुसार(

भवदीय/भवदीया/आपका

)नाम (

अनुलग्नक उपरोक्त :अथवा उपरोक्त वर्णित

किसी अन्य माध्यम से प्राप्त पत्र का उत्तर देने के लिए हार्ड कॉपी के) - (रूप में प्राप्त पत्र का उत्तर देने के लिए भी

महोदय/महोदया/प्रिय ..... आदरणीय/,

उपरोक्त विषय के संबंध में आपके दिनांक के संदर्भ में ..... के पत्र संख्या .....,

- कृपया संलग्न पत्र का अवलोकन करें ।
- इस संस्थान से संबन्धित जानकारी/सूचना संलग्न है।
- संस्थान के निदेशक का पत्र संलग्न है।

सधन्यवाद । / साभार ।)आवश्यकतानुसार(

भवदीय/भवदीया/आपका

अपनी ओर से ई- मेल के माध्यम से पत्र भेजने के लिए-

विषय :

महोदय/ महोदया/प्रिय ..... आदरणीय/

- कृपया उपरोक्त विषय के संबंध में इस मेल के साथ संलग्न पत्र का अवलोकन करें।
- उपरोक्त विषय के संबंध में कृपया अनुलग्नक का अवलोकन करें।
- उपरोक्त विषय के संबंध में निदेशक की ओर से जारी संलग्न पत्र का अवलोकन करने की कृपा करें।
- उपरोक्त विषय के संबंध में निदेशक की ओर से जारी पत्र/कार्यालय आदेश परिपत्र/संलग्न है ।
- उपरोक्त विषय के संबंध में निदेशक की ओर से जारी संलग्न पत्र/कार्यालय आदेश परिपत्र/ आवश्यक कार्रवाई हेतु प्राप्त करें ।

ई-मेल के माध्यम से पृष्ठांकन जारी करने के लिए -

विषय : ...

उपरोक्त विषय के संबंध में परिषद से प्राप्त दिनांक /का पत्र 2018 दिसंबर 15 कार्यालय आदेश परिपत्र/सं॰ ..... आवश्यक कार्रवाई हेतु पृष्ठांकित/अनुपालन हेतु संलग्न है/सूचनार्थ/ " किया जा रहा है । कृपयाकी"गई कार्रवाई रिपोर्ट- शीघ्र उपलब्ध कराएं ।

दूरभाष पर हुई चर्चा के संदर्भ में -

विषय ; ...

- उपरोक्त विषय के संदर्भ में हाल ही में आपके साथ हुई चर्चा के दृष्टिगत, अपेक्षित जानकारी सूचना संलग्न है। कृपया आवश्यक कार्रवाई हेतु देखें।/
- उपरोक्त विषय के संदर्भ में हाल ही में आपके साथ हुई चर्चा के दृष्टिगत, अपेक्षित आंकड़े प्रेषित किए जा रहे हैं, कृपया आवश्यक कार्रवाई हेतु प्राप्त करें तथा पावती की सूचना दें ।

समस्त सामग्री यूनिकोड में टंकित है । ई मेल में हिन्दी यूनिकोड का ही उपयोग किया जाए- ।

नोट: कार्यालय आदेश, ज्ञापन, परिपत्र, नोटिस, इत्यादि धारा (3)3के अंतर्गत आने वाले सभी कागजात द्विभाषी जारी होने आवश्यक हैं ।

कृषि में जैविक डेटा विश्लेषण के लिये जैव कंप्यूटिंग पोर्टल, टूल्स और एल्गोरिदम

प्रस्तावना

भारत में कृषि विभिन्न भौतिक, रसायनविज्ञानी एवं जीवविज्ञानी घटकों का सम्मिलन है जो जब एक विशेष पद्धति में परस्पर जुड़ते हैं तो उसके परिणामस्वरूप देश की बढ़ती हुई जनसंख्या को भोजन उपलब्ध कराने के लिए फसलों की उत्पादकता बढ़ाना संभव होता है। वर्तमान कृषि *हरित क्रांति* से *सदाबहार हरित क्रांति* की ओर तेजी से अग्रसर हो रही है जो न केवल मानव सभ्यता के लिए लाभकारी है बल्कि कृषि की देसी प्राकृतिक जैविक युक्तियों, मृदा व जल के हेतुविज्ञान तथा अन्य आवासों के लिए भी लाभप्रद है। इसके साथ ही *छोटे किसानों के लिए चतुराईपूर्ण खेती* की भी आवश्यकता है।

भारत में पिछले दो दशकों के दौरान कृषि जैवप्रौद्योगिकी, आण्विक जैव–प्रौद्योगिकी, जीनोमिक्स तथा सम्बद्ध विज्ञानों के क्षेत्र में अनुसंधान प्रयास किए जा रहे हैं लेकिन हमारे देश की अधिकांश जनसंख्या अब भी इन प्रयासों से होने वाले उल्लेखनीय व वांछित परिणामों की प्रतीक्षा कर रही है। अब कृषि विज्ञानों से संबंधित सभी जीवविज्ञानी विषयों में आंकड़ा–आधारित, बहु–ओमिक्स अनुसंधान अनिवार्य है। वर्तमान में फसल पौधों, पशुधन, सूक्ष्मजीवों, कीटों, हैबिटेट तथा अन्य संबंधित विषय क्षेत्रों के गुणप्ररूपी व जीनप्ररूपी जीवविज्ञानी आंकड़े अत्यधिक तीव्र गति से सृजित किए जा रहे हैं – ओमिक्स अनुसंधान (जीनोमिक्स, प्रोटियोमिक्स, मेटाबोलोमिक्स आदि), कम्प्यूटेशनल जीवविज्ञान व जैव सूचना विज्ञान से सृजित इन वृहत आंकड़ों से जीवविज्ञानी प्रक्रिया को समझने के लिए सूचना एवं ज्ञान प्राप्त करना पूरे विश्व के लिए आवश्यक हो गया है। इससे आंकड़ों के विश्लेषण, पूर्वानुमान, भंडारण, प्रबंध, पैटर्न को पहचानने, प्रस्तुतीकरण, पुनःप्राप्ति और भंडारण की समस्याओं को हल करने में सहायता मिलती है ताकि इनका एक सार्थक मूल्य ज्ञात किया जा सके। पौधों, पशुधन, कीटों, मछलियों व सूक्ष्मजीवों पर पूर्ण हो चुकी या चल रही सम्पूर्ण जीनोम अनुक्रमण परियोजनाओं से विशाल सूचना विज्ञानी संसाधन सृजित हुए हैं जिनसे विभिन्न विषयों के वैज्ञानिकों/अनुसंधानकर्ताओं को एक–दूसरे के साथ मिल–जुलकर कार्य करने के लिए मजबूर होना पड़ा है। कृषि की दृष्टि से महत्वपूर्ण गुणों का पता लगाने के लिए कोशिकाओं में छिपे हुए ज्ञान का गहराई से अवलोकन करने, जीवों में या उनके बीच की अंतरक्रियाओं तथा पर्यावरण में जैविक व अजैविक प्रतिबलों की अनुक्रियाओं को समझने, इन्हें उनके जीनों, विशेषकों आदि से जोड़ने, प्रोटीनों से सम्बद्ध करने, मूल्यवर्धित उपापचयी उत्पादों को तैयार करने, गुणप्ररूपी विशेषताओं को समझने, अनुकूलन व्यवहार, विकासात्मक पैटर्न व संभावित अंतरक्रियाओं को जानने के लिए जीवविज्ञान के विषयों के वैज्ञानिक कम्प्यूटेशनल वैज्ञानिकों के साथ दल भावना के साथ कार्य करने के लिए बाध्य हुए हैं। बहु–ओमिक्स जीवविज्ञानी युग में जीव के प्रत्येक विशेषक और गुण का –ओम से संबंध है जैसे फीनोम, जीनोम, इपिजीनोम, ट्रांसक्रिप्टोम, प्रोटियोम, मेटाबोलोम, रिएक्टोम, लोकलाइजोम, इंटरेक्टोम, राइजोम, बायोम, माइक्रोबायोम, ट्राइकोम, वैक्योम आदि, ताकि कार्यों को उनकी पूर्णता व सम्पूर्णता में व्यक्त किया जा सके। जीवों के प्रत्येक पहलू पर सृजित आंकड़ों का आकार इतना बड़ा है कि उसकी व्याख्या करना, उसका विश्लेषण, भंडारण, प्रबंध व उसे पुनः प्राप्त करना जीवविज्ञानी पैटर्न पर उल्लेखनीय है। अतः वर्तमान कृषि अनुसंधान एवं शिक्षा में

जैवसूचना विज्ञान तथा उसके एकीकरण की आवश्यकता है ताकि इसे कम्प्यूटर के उच्च अनुप्रयोगों, युक्तियों व सॉफ्टवेयर, डेटाबेस के विकास एवं प्रबंध, कम्प्यूटेशनल जीवविज्ञान, जैवप्रौद्योगिकी तथा जैवसांख्यिकी में उपयोगी ढंग से इस्तेमाल किया जा सके। इस क्षेत्र में होने वाला अनुसंधान बहु या अंतरविषयी प्रकृति का है क्योंकि किसी एक विषय से इन जीवविज्ञानी प्रक्रियाओं के रहस्य को समझा नहीं जा सकता है। किसानों के लिए प्रभावी जैवप्रौद्योगिकी उत्पाद/जिंसें विकसित करने के लिए कृषि विज्ञान के लगभग सभी विषयों को जिनमें प्रजनन, सस्यविज्ञान, पादप सुरक्षा आदि भी शामिल हैं, को एकीकृत करने की आवश्यकता है।

कम्प्यूटेशनल युक्तियों से संबंधित विकास ने वैश्विक स्तर पर इस दिशा में किए जाने वाले प्रयासों को परस्पर जोड़ दिया है तथा पिछले एक दशक के दौरान जीवविज्ञान संबंधी अनुसंधान में क्रांतिकारी परिवर्तन हुए हैं। वर्तमान में जीवविज्ञानी वैश्विक स्तर पर अनेक विषयों के वैज्ञानिकों के साथ मिल–जुलकर कार्य कर रहे हैं ताकि जटिल जीवविज्ञानी प्रणालियों के कार्यों को उजागर किया जा सके। आनुवंशिक अभियांत्रिकी तथा जीनोमी युक्तियों ने जैव प्रणालियों की उत्पादकता व गुणवत्ता संबंधी गुणों को बढ़ाने के लिए नए द्वार खोले हैं। जीनोमिक डेटाबेस में बड़ी मात्रा में ऐसी सूचना होती है जिसे विश्लेषण की परंपरागत युक्तियों से प्राप्त करना संभव नहीं है। इसलिए जैव सूचना विज्ञान तथा कम्प्यूटेशनल जीवविज्ञान एक अंतरविषयी कार्यक्रम के रूप में उभरे हैं जिसमें कम्प्यूटेशनल तथा गणितीय विज्ञानों को जीवन से जुड़े विज्ञानों के साथ जोड़ा गया है। कम्प्यूटेशनल जीवविज्ञान तथा कृषि जीवसूचना विज्ञान का उद्देश्य जीववैज्ञानिकों, सांख्यिकीविदों तथा कम्प्यूटर वैज्ञानिकों को एक साथ जोड़ना है। पिछले दो दशकों के दौरान कृषि जैवप्रौद्योगिकी अनुसंधान में गहन प्रयास हुए हैं लेकिन इन प्रयासों का वास्तविक प्रभाव फार्म स्तर पर अब भी प्रतीक्षित है। इसका मुख्य कारण हमारे देश में जीवविज्ञानी कम्प्यूटिंग से संबंधित बुनियादी ढांचे का न होना है। सांख्यिकी तथा कम्प्यूटेशनल विज्ञानों का उपयोग करके जीनोमिक्स सूचना तथा मानव ज्ञान के बीच के अंतराल को कम करने के लिए कम्प्यूटेशनल संबंधी बुनियादी ढांचे की आवश्यकता है। बड़े जीनोमी डेटाबेस, डेटा वेयरहाउस, सॉफ्टवेयर और तकनीकियां, एल्गोरिथ्म व उच्च स्तर की कम्प्यूटेशनल शक्ति वाले जीनोम ब्राउजरों के लिए इनकी आवश्यकता है ताकि संबंधित प्रजातियों के जीनोमी संसाधनों से सूचना और ज्ञान को प्राप्त किया जा सके। जैव सूचना विज्ञान में डाउन स्ट्रीम अनुसंधान के क्षेत्र में नए आयाम खोलने के लिए भी बुनियादी ढांचे की आवश्यकता है। इसके अंतर्गत कोशिकीय कार्यों की मॉडलिंग, आनुवंशिक नेटवर्क, मेटाबॉलिक (उपापचय) पथों, जीन के कार्यों को समझने के लिए लक्ष्यों के सत्यापन और उन्नत किस्मों तथा नस्लों के विकास के लिए जीनों के कार्यों को समझने व उनका उपयोग करने के लिए भी अपेक्षित स्तर के बुनियादी ढांचे की जरूरत है ताकि कृषि उत्पादकता को कई गुना बढ़ाया जा सके। किस्मों/नस्लों के विकास के द्वारा जलवायु परिवर्तन की चुनौतियों से निपटने व मानवीय स्वास्थ्य संबंधी समस्याओं को हल करने के लिए (प) जैविक (रोगों) तथा लवणता, तापमान, सूखा आदि जैसे अजैविक प्रतिबलों से निपटना, (पप) कम कार्बन डाइऑक्साइड उत्सर्जित करना व (पपप) आर्सेनिक, नाशकजीवनाशियों, कृषि रसायनों जैसे पर्यावरणीय प्रदूषकों के अपशिष्ट को कम करना जैसी समस्याओं से निपटना भी महत्वपूर्ण है जो देश में कृषि जैवप्रौद्योगिकी अनुसंधान के द्वारा ही संभव है। इससे न केवल वैश्विक स्तर पर ऐसे कृषि उत्पादों को तैयार करने में सहायता मिलेगी जिनसे हमारा कृषि निर्यात बढ़ सकता है बल्कि हमारे बौद्धिक सम्पदा अधिकारों व

पेटेंटों की भी रक्षा होगी। इससे अगली सदाबहार हरित क्रांति आएगी जिसके द्वारा देश में पोषणिक खाद्य एवं आजीविका सुरक्षा सुनिश्चित होगी। अनेक जीवों के जीनोम अनुक्रमण से कई रोचक तथ्य उजागर हुए हैं। आज संसाधनों और समय की बचत के लिए इस विषय की आवश्यकता विश्वभर में अनुभव की जा रही है।

कृषि एवं जीवविज्ञान में जैवसूचना विज्ञान एवं कम्प्यूटेशनल जीवविज्ञान का मूल कार्य उन जीवविज्ञानी प्रश्नों के उत्तर खोजना है जिनका उपयोग कृषि उत्पादन और उत्पादकता को सुधारने तथा कृषि पारिस्थितिक प्रणाली को टिकाऊ बनाए रखने में किया जा सकता है। इस क्षेत्र में अनिवार्य रूप से तीन मौलिक कार्य घटक हैं :

- वृहत जीवविज्ञानी डेटासेटों के भंडारण व प्रबंध के लिए डेटाबेसों का सृजन;

- नए एल्गोरिथ्म व सांख्यिकी कार्यक्रमों, गणितीय अनुरूपण माडलों, मशीन द्वारा सीखने की तकनीकों, हाई–एंड सॉफ्टवेयर, कस्टमाइज्ड कम्प्यूटर कार्यक्रमों व बड़े डेटा सेटों की इकाइयों के बीच के संबंध का पता लगाने के लिए भाषाओं का विकास; तथा

- विभिन्न प्रकार की जीवविज्ञानी आंकड़ा तकनीकों के बीच की अंतरक्रियाओं के साथ–साथ आंकड़ा समेकन, भागेदारी, आंकड़ों की व्याख्या व विश्लेषण के लिए युक्तियों का उपयेाग।

एक प्रकार से जैवसूचना विज्ञान में वृहत, विविध तथा जटिल जीवविज्ञानी आंकड़ों को और अधिक पठनीय, समझने योग्य व उपयोग योग्य बनाने के लिए सूचना विज्ञान संबंधी प्रौद्योगिकियों के सिद्धांतों को लागू किया जाता है, जबकि कम्प्यूटेशनल जीवविज्ञान में प्रयोगात्मक तथा सैद्धांतिक शंकाओं के समाधान के लिए एल्गोरिद्म, गणितीय मॉडलों व कम्प्यूटेशनल युक्तियों का उपयोग किया जाता है। इस प्रकार, कार्यों तथा दृष्टिकोणों में भिन्न होने के बावजूद भी इनकी क्रियाओं में उल्लेखनीय समानता है जिससे जीवविज्ञान के किसी भी विषय को सूचना विज्ञान के साथ जोड़कर उपयोग में लाया जा सकता है।

इस क्षेत्र का एक मुख्य उद्देश्य प्रणाली जीवविज्ञान दृष्टिकोण व समस्या को प्रभावी ढंग से हल करने के लिए जीव वैज्ञानिकों, सांख्यिकीविदों व कम्प्यूटर वैज्ञानिकों को एक साथ लाना है। इस दृष्टिकोण से राष्ट्रीय व अंतरराष्ट्रीय संगठनों के बीच विभिन्न स्तरों पर साझेदारियां विकसित करने में सहायता प्राप्त होती है। कृषि सूचना विज्ञान, कम्प्यूटेशनल जीवविज्ञान व संबंधित क्षेत्रों में अनुसंधानकर्ताओं व वैज्ञानिकों के बीच कार्यात्मक सम्पर्क स्थापित किए जाने की आवश्यकता है। छोटे पैमाने पर भा.कृ.अ.प. के विभिन्न संस्थानों में आरंभ किए गए जैवसूचना विज्ञान से संबंधित क्रियाकलापों का हाल ही में उन्नयन किया गया है तथा कृषि के क्षेत्र में इन्हें राष्ट्रीय स्तर पर प्रोत्साहित भी किया गया है। देसी कृषि जीनोमी संसाधनों के संकलन, संचयन, भंडारण तथा संबंधित ज्ञान के खनन के वृहत प्रयास किए गए हैं। कृषि जैव सूचना विज्ञान में अनुसंधान व विकास के मामले में वैश्विक स्तर पर हुई प्रगति के समान प्रगति करने के लिए देश को *राष्ट्रीय कृषि जैव सूचना* विज्ञान ग्रिड के माध्यम से अनुसंधान के क्षेत्र में और अधिक विशेषज्ञता प्राप्त करने व विश्व में उपलब्ध ज्ञान का उपयोग करने की आवश्यकता है। यह ग्रिड भारतीय कृषि अनुसंधान परिषद, नई दिल्ली द्वारा स्थापित किया गया है जिसमें संबंधित डेटाबेस, डेटा वेयरहाउस, सॉफ़टवेयर और युक्तियां, एल्गोरिद्म, जीनोम ब्राउजर आदि विकसित किए जा रहे हैं। भारतीय कृषि के लिए प्रथम सुपर कम्प्यूटर

(अशोका) के रूप में उच्च स्तर की कम्प्यूटिंग सुविधा कृषि अनुसंधानकर्ताओं के लिए उपलब्ध कराई गई है जिसके लिए क्रमबद्ध तथा एकीकृत दृष्टिकोण का उपयोग किया गया है। इस सुविधा के द्वारा भा.कृ.अ.प. के विभिन्न संस्थानों के विभिन्न जीनोमिक आंकड़ों का विश्लेषण किया जा रहा है। दीर्घावधि में जीनोमी ज्ञान आधार से अंतर–विषयी अनुसंधान के माध्यम से सृजित सूचना और ज्ञान का प्रवाह नीचे की ओर होगा तथा कृषि के विभिन्न क्षेत्रों में किए जा रहे प्रयोगों में इसका लाभ उठाया जा सकेगा जिससे अंतरराष्ट्रीय स्तर पर श्रेष्ठ प्रतिस्पर्धी किस्में/नस्लें व कृषि से जुड़ी जिंसें विकसित करना संभव होगा। इसके साथ ही भारतीय कृषि अनुसंधान परिषद, नई दिल्ली की *'राष्ट्रीय कृषि नवोन्मेष परियोजना (एनएआईपी)'* के अंतर्गत *'राष्ट्रीय कृषि जैवसूचना विज्ञान ग्रिड (एनएबीजी)'* पर उप परियोजना से विभिन्न प्रजातियों की जीनेामिक्स में बहु–विषयी अनुसंधान के लिए मंच उपलब्ध हुआ है जिससे अनेक सहयोत्मक अनुसंधान परियोजनाएं तैयार हुई हैं तथा भा.कृ.अ.प. के विभिन्न संस्थानों में गुणवत्तापूर्ण प्रकाशन निकाले गए हैं।

अशोकाः भारतीय कृषि अनुसंधान के लिए प्रथम सुपर कम्प्यूटिंग हब (कृषि अनुसंधान में मील का पत्थर)

कृषि जैव सूचना विज्ञान केन्द्र, भारतीय कृषि अनुसंधान संस्थान, नई दिल्ली, भारत में भारतीय कृषि के लिए प्रथम सुपर कम्प्यूटिंग हब *अशोका* (एडवांस कम्प्यूटिंग हब फॉर ओमिक्स नॉलेज इन एग्रीकल्चर) स्थापित किया गया है। स्थापित की गई यह सुविधा एक अति उत्कृष्ट आंकड़ा केन्द्र है तथा इस हब के दो सुपर कम्प्यूटर (http://topsupercomputers-india.iisc.ernet.in/jsps/june2013/index.html) में भारत के सर्वश्रेष्ठ सुपर कम्प्यूटरों की सूची में 11वें और 24वें स्थान पर है।

इस सुपर कम्प्यूटिंग हब में (i) 3072 कोर और 38 टेरा प्लॉक्स कम्प्यूटिंग से युक्त दो मास्टर तथा 256 नोड लीनक्स क्लस्टर, (ii) एक मास्टर कम्प्यूटर से युक्त 16 नोड वाला विंडोज (iii) 192 सीपीयू + 8192 (जीपीयू) से युक्त एक मास्टर नोड तथा 16 नोड्स जीपीयू क्लस्टर तथा (iv) 1.5 टीबी रैम से युक्त एस.एम.पी. आधारित मशीन है। इस हब में लगभग 1.5 पेटा बाइट भंडारण है जिसे तीन विभिन्न प्रकार की भंडारण आर्किटैक्चर अर्थात् नेटवर्क एटैच्ड स्टोरेज (एन.ए.एस.), पैरलल फाइल सिस्टम (पी.एफ.एस.) और आर्काइवल में बांटा गया है। इस हब में अन्य सुपर कम्युनिटिंग प्रणालियां भी हैं (40 टीबी भंडारण से युक्त एक मास्टर नोड वाला 16 नोड लीनक्स क्लस्टर) जो भा.कृ.अ.प. – राष्ट्रीय पादप आनुवंशिक संसाधन ब्यूरो (एन.बी.पी.जी.आर.), नई दिल्ली; भा.कृ.अ.प.– राष्ट्रीय पशु आनुवंशिक संसाधन ब्यूरो (एन. बी.ए.जी.आर.), करनाल; भा.कृ.अ.प.– राष्ट्रीय मत्स्य आनुवंशिक संसाधन ब्यूरो (एन.बी.एफ.जी. आर.) लखनऊ; भा.कृ.अ.प.– कृषि की दृष्टि से महत्वपूर्ण जीवों के राष्ट्रीय ब्यूरो (एनबीएआईएम), मऊ; तथा भा.कृ.अ.प.– राष्ट्रीय कृषि कीट संसाधन ब्यूरो (एन.बी.ए.आई.आर.), बंगलुरू में स्थित है तथा देश में राष्ट्रीय कृषि जैव सूचना विज्ञान ग्रिड का निर्माण करता है। राष्ट्रीय जीवविज्ञानी कम्प्यूटिंग पोर्टल के साथ–साथ अनेक कम्प्यूटेशनल जीवविज्ञान व कृषि जैवसूचना विज्ञान संबंधी सॉफ्टवेयर/ वर्कफ्लो/ पाइपलाइन विकसित किए गए हैं जो देशभर में जीवविज्ञानी अनुसंधानकर्ताओं को जीवविज्ञानी कम्प्यूटिंग संसाधन उपलब्ध कराते हैं।

जीवविज्ञान डेटाबेसों का विकास

कृषि से संबंधित जीवविज्ञानी डेटाबेसों का विकास कृषि जैवसूचना विज्ञान के क्षेत्र में बहुत महत्वपूर्ण है। टमाटर के माइक्रोसेटेलाइट डीएनए मार्कर डेटाबेस पर आधारित प्रथम सम्पूर्ण जीनोम, टोमेटो माइक्रो सेटेलाइट डेटाबेस (TomSatDb), http://webabp.cabgrid.res.in/tomsardb/ में कुल 1466002 एस.टी.आर. मार्कर हैं (इकबाल और साथी 2013)। वैट लैब, की आवश्यकताओं को पूरा करने के लिए स्वचालित प्राइमर डिजाइनिंग तकनीक भी जोड़ी गई है। TomSatDB उपयोगकर्ता–मित्र है तथा इसकी मुक्त पहुंचने योग्य तकनीक प्राइमरों की स्थलवार व गुणसूत्रवार खोज का अवसर प्रदान करती है। इन मार्करों से अजैविक व जैविक प्रतिबलों के लिए जननद्रव्य के प्रबंध का मार्ग प्रशस्त होने के साथ–साथ आनुवंशिक प्रजनन के माध्यम से किस्मों के सुधार की भी आशा है जिससे विश्व के विभिन्न भागों में टमाटर की उत्पादकता में वृद्धि होगी। अजैविक प्रतिबल के अतिरिक्त टमाटर की फसल में रोगजनक कवकों, जीवाणुओं, विषाणुओं तथा सूत्रकृमियों द्वारा 200 से अधिक रोग लगते हैं जो टमाटर की उत्पादकता को प्रतिकूल रूप से प्रभावित करते हैं। वांछित उत्पादकता के साथ–साथ अजैविक व जैविक प्रतिबल के लिए जननद्रव्य के प्रबंध हेतु ऐसे घनिष्ठ रूप से संबंधित डीएनए मार्करों की अत्यधिक आवश्यकता है। इसके अलावा विशेष रूप से नई किस्मों के विकास संबंधी कार्यक्रम में मार्कर सहायी समाहन हेतु आर्थिक व वाणिज्यिक रूप से महत्वपूर्ण जीनों का उपयोग किया जा सकता है। ये निष्कर्ष टमाटर जीनोमिक्स अनुसंधान में अत्यधिक उपयोगी सिद्ध हो सकते हैं तथा वैश्विक स्तर पर टमाटर के सुधार व किस्म प्रबंध के प्रयास में भी कारगर हो सकते हैं। इसी प्रकार, अरहर के लिए माइक्रो सेटेलाइट डेटाबेस (PIPEMicroDB), http://webapp.cabgrid.res.in/pigeonpea/ (सारिका और साथी 2013) भी विकसित किए गए हैं। ये मार्कर, मार्कर सहायी चयन में अत्यधिक उपयोगी सिद्ध होंगे जिनसे भारत में तथा विश्व के अनेक भागों में जैविक व अजैविक प्रतिबलों के कारण अरहर की उत्पादकता में होने वाली लगभग 50 प्रतिशत कमी को दूर किया जा सकेगा। इसके अतिरिक्त 910529 माइक्रो सेटेलाइट मार्कर http://webapp.cabgrid.res.in/buffsatdb/ (सारिका और साथी, 2013) से युक्त भैंस माइक्रोसेटेलाइट डेटाबेस भी अंतरराष्ट्रीय समुदाय को उपलब्ध कराया गया है। इस डेटाबेस को चुने हुए मार्करों की प्राइमर डिजाइनिंग के लिए प्राइमर–3 से जोड़ा गया है जिससे अनुसंधानकर्ताओं द्वारा गुणसूत्रों को वांछित अंतराल पर चुनने में सहायता प्राप्त होगी। डिजेनरेट आधारों के पुनः जुड़ने से वर्तमान भैंस जीनोम एसेम्बली में डिजेनरेट आधारों की उपस्थिति की समस्या को हल करने में भी सहायता मिलेगी। विश्व में भैंस का प्रथम एस.टी. आर. डेटाबेस होने के कारण इससे न केवल वर्तमान एसेम्बली संबंधी समस्या को सुलझाने का मार्ग प्रशस्त होगा बल्कि यह क्यूटीएल/जीन मानचित्रण में वैश्विक समुदाय के लिए अत्यधिक उपयोगी सिद्ध होगा जिसकी विशेष रूप से तृतीय विश्व के देशों में भैंसों की उत्पादकता को बढ़ाने के लिए बहुत आवश्यकता है क्योंकि यहां की ग्रामीण अर्थव्यवस्था भैंसों की उत्पादकता पर ही मुख्यतः निर्भर है। इन मार्करों का उपयोग जनकता के परीक्षण, नस्ल की पहचान, समष्टि की संरचना तैयार करने तथा अपमिश्रण के विश्लेषण हेतु किया जा सकता है। इनका उपयोग विशेष रूप से जननद्रव्य विनिमय या जननद्रव्य के सीमा पार संबंधी आने–जाने से जुड़े मुद्दों के लिए भी किया जा सकता है। एक बकरी माइक्रोसेटेलाइट डेटाबेस (GoSatDb), http://webapp.cabgrid.res.in/goat भी विकसित किया गया है जिसमें बकरी के सम्पूर्ण जीनोम क्रम में 865210 माइक्रो सेटेलाइट मार्कर मौजूद हैं।

लवण रागी आर्की/जीवाणु लवण की विभिन्न सांद्रताओं अर्थात् अत्यधिक, मध्यम और निम्न के प्रति अपने को अनुकूल ढाल लेते हैं। इस प्रकार के अनुकूलन विभिन्न कोशिका उपांगों में होने वाले अन्य परिवर्तनों व प्रोटीन संरचना में रूपांतरण के परिणामस्वरूप होते हैं। इस प्रकार प्रोटीनें लवण रागी आर्की/जीवाणुओं के लवणीय स्थितियों के प्रति अनुकूलन में महत्वपूर्ण भूमिका अदा करते हैं। लवण रागी प्रोटीन डेटाबेस (HPortDB) लवणरागी आर्की/जीवाणुओं से प्राप्त प्रोटीनों के जैव रसायनविज्ञानी व जैव भौतिकी गुणों के प्रलेखन का एक क्रमबद्ध प्रयास है जो इन जीवों के लवणीय स्थितियों के प्रति अनुकूलन में शामिल हो सकता है। इस डेटाबेस में विभिन्न भौतिक–रसायनविज्ञानी गुणों जैसे आण्विक भार, सैद्धांतिक च, एमिनो अम्ल संघटन, परमाण्विक संघटन, अनुमानित अर्ध जीवन, अस्थिरता सूचकांक, एलिफेटिक सूचकांक तथा महा औसत हाइड्रोपैथिसिटी (ग्रैवी) को सूचीबद्ध किया गया है। ये भौतिक–रसायनविज्ञानी गुण प्रोटीन संरचना, उनके बंधन पैटर्न व विशिष्ट प्रोटीनों के कार्यों को पहचानने में महत्वपूर्ण भूमिका निभाते हैं। यह डेटाबेस वृहत है, इसे मानवीय रूप से क्यूरेट किया जा सकता है तथा यह प्रोटीनों का नॉन–रिडंडेंट केटालॉग है। इस डेटाबेस में वर्तमान में 59 897 प्रोटीन गुण हैं जो विभिन्न प्रकार के लवण रागी आर्की/जीवाणुओं के 21 प्रभेदों से निष्कर्षित किए गए हैं। इस डेटाबेस तक वेबसाइट के लिंक के माध्यम से पहुंचा जा सकता है। डेटाबेस URL:http://webapp.cabgrid.res.in/protein/ (नवीन और साथी, 2014)।

इपिजेनेटिक्स का अर्थ जीन अभिव्यक्ति में उन परिवर्तनों से है जो डीएनए क्रम के परिवर्तन में शामिल नहीं है। इस संकल्पना का अर्थ है कि एक बार स्थापित हो जाने पर नई आनुवंशिक अवस्था को प्रजननशील संकेत के माइटोसिस या मियोसिस से स्वतंत्र रखते हुए स्थायी रूप से प्रवर्धित किया जा सकता है और उसके बाद भी उसे अपनी मूल अवस्था में वापस लाया जा सकता है। पशुधन प्रजातियों में इपिजेनेटिक यांत्रिकियों से संबंधित सूचना एक स्थान पर उपलब्ध नहीं है। इसके अलावा पशुधन में उत्पादन संबंधी गुणों को सुधारने तथा रोगों के नियंत्रण के लिए इपिजेनेटिक सूचना के विश्लेषण की आवश्यकता होती है। एक वैब आधारित 'पशुधन इपिजेनेटिक सूचना प्रणाली' विकसित की गई है (http://bioinformatics.iasri.res.in/edil/) जिसमें बॉटम लेयर के रूप में MySQL डेटाबेस है, सर्वर साइट एप्लीकेशन–मिडल लेयर के रूप में पीएचपी तथा टॉप लेयर पर एचटीएमएल, सीएसएस और जावा स्क्रिप्ट हैं।

पशु जीनोम पर भैंस जीनों के ऑर्थोलॉगस एनोटेड डेटा प्राप्त करने के लिए php स्क्रिप्ट व MySQL डेटाबेस का उपयोग करके एक वैब इंटरफेस विकसित किया गया है। गोपशु जीनोम पर भैंस के जीनों के मानचित्रण के लिए लाइट वेट जीनोम व्यूवर तकनीक का उपयोग करके एक ब्राउजर विकसित किया गया है। जीनोम की रचना के लिए वांछित एनोटेशन फाइलें भी तैयार की गई हैं। भैंस के जीनोम के विभिन्न कार्यात्मक तत्वों पर सूचना पोपुलेट की गई है। भैंसों के जीनोम पर इन तत्वों का मानचित्रण किया गया है। मानचित्रित सूचना को लाइटवेट जीनोम ब्राउजर तकनीक द्वारा प्रदर्शित किया गया है । भैंस जीनोम डेटाबेस व ब्राउजर तकनीक सहित एक वेबसाइट तैयार की गई है। गोपशु तथा भैंस जीनोमों के साथ भैंस के जीन संबंधी सूचना का एकीकरण व मानचित्रण भी उपयोगकर्ताओं के लिए किया गया है।

जैव–कम्प्यूटिंग पोर्टल एवं तकनीकों का विकास

राष्ट्रीय कृषि जैव कम्प्यूटिंग पोर्टल (http://webapp.cabgrid.res.in/biocomp/) उच्च निष्पादन कम्प्यूटिंग (एचपीसी) संसाधनों तक पहुंच के लिए एकल बिंदु उपलब्ध कराता है। यह पोर्टल जैव सूचना विज्ञानी कार्यों को सम्पन्न करने के लिए एक वातावरण उपलब्ध कराता है। यह पोर्टल उपयोगकर्ता को विशिष्ट जॉब प्रस्तुत करने व प्रबंध संबंधी अनुप्रयोगों को सम्पन्न करने में सहायता प्रदान करता है। इस पोर्टल द्वारा प्रस्तुत किए गए कार्य अनुसूचित होते हैं और संसाधनों का आबंटन रिसोर्स मैनेजर के माध्यम से किया जाता है। यह रिसोर्स मैनेजर, संसाधनों तक पहुंच, उनके आबंटन तथा प्रबंध व कार्यों के कार्यान्वयन से संबंधित है। उपयोगकर्ता पोर्टल इंटरफेसों के माध्यम से इनपुट और आउटपुट डेटा को प्रबंधित कर सकता है। उपयोगकर्ताओं को यूजर्स लॉगिंग का उपयोग करके पोर्टल में लॉग–इन करना होता है, ताकि पोर्टल की सुविधाओं जैसे जॉब्स प्रस्तुतीकरण, जॉब स्तर की ट्रैकिंग और व्यू हाउटपुट/इरर डेटा तक पहुंचा जा सके। यह विशिष्ट प्राचलों के साथ कार्यों की निगरानी तथा प्रस्तुतीकरण के लिए वैब इंटरफेस उपलब्ध कराता है। क्रमबद्ध तथा समानांतर एप्लिकेशंस जॉब को पोर्टल के माध्यम से प्रस्तुत किया जा सकता है। यह ग्रिड उपयोगकर्ताओं को ऐसा पर्यावरण उपलब्ध कराती है जिसमें कम्प्यूटिंग की अदम्य शक्ति होती है, बड़ी मात्रा में आंकड़े होते हैं तथा बड़ी संख्या में प्रणालियां या संसाधन होते हैं। संसाधन निर्देशिकाओं या रिसोर्स डायरैक्टरीज का उपयोग ग्रिड संसाधनों की स्थिति, संरचना और अवस्था के बारे में सूचना प्राप्त करने के लिए किया जाता है। इस सुविधा के परिचालनात्मक प्रबंध से संबंधित मुद्दों से निपटने के लिए चौबीसों घंटे उपलब्ध हैल्प डैस्क सहायता भी मौजूद है। इन कम्प्यूटेशनल संसाधनों के प्रबंध के लिए स्वचालित तकनीकों के विभिन्न सैट कान्फीग्यूरेट किए गए हैं। इससे देश में जैवप्रौद्योगिकी अनुसंधान की कम्प्यूटेशन से संबंधित आवश्यकताओं की पूर्ति होगी। सांख्यिकीय तथा कम्प्यूटेशनल विज्ञानों का उपयोग करके इसके द्वारा जीनोमी सूचना और ज्ञान के बीच मौजूदा अंतराल को भी पाटा जा सकेगा। इससे बड़े जीनोमी डेटाबेस, डेटा वेयरहाउस, सॉफ्टवेयर तकनीकियां एल्गोरिथ्म तथा उच्च एंड कम्प्यूटेशनल शक्ति वाले जीनोमी ब्राउजर स्थापित करने में सहायता मिलेगी जिससे पार–प्रजाति जीनोम संसाधनों से सूचना व ज्ञान प्राप्त किए जा सकेंगे।

क्रम प्रस्तुतीकरण प्रोर्टल : पौधों, पशुओं, कीटों, सूक्ष्मजीवों तथा मात्स्यिकी की जीवविज्ञान सूचना से संबंधित वृहत आंकड़े सृजित करने के लिए कृषि वैज्ञानिकों द्वारा अनेक अध्ययन किए जाते हैं। ये एनसीबीआई, ईएमबीएल, डीडीबीजे तथा अन्य पोर्टलों पर अपने जीव आंकड़े प्रस्तुतीकरणों के लिए निर्भर रहते हैं। इन स्थलों द्वारा लगाए गए विभिन्न प्रतिबंधों तथा घटिया कनेक्टिवीटी की समस्या के कारण इन खुले डोमेन के डेटाबेसों पर अध्ययन करना संभव नहीं हो पाता है। अतः एक सुरक्षित क्रम प्रस्तुतीकरण पोर्टल विकसित किया गया है जिसमें बैकएंड डेटाबेस होता है और मानक डेटाबेस प्रबंध संकल्पनाओं को अपनाया जाता है (http://webapp.cabgri.res.in/denadb/)(लाल और साथी)। यह पहल देश में देसी जीनोमी डेटाबेस का निर्माण करने व विश्लेषण मंच तैयार करने के लिए की गई है। इस डेटाबेस से उच्च गति के सूचना संसाधन व ज्ञान को प्राप्त करने के लिए प्रगत हार्डवेयर संसाधन व समानांतर कम्प्यूटिंग सुविधाएं विकसित की गई हैं। डिजाइन किए गए डेटाबेस को जैवसूचना विज्ञान के क्षेत्र में कार्यरत कृषि वैज्ञानिकों द्वारा विकसित विशाल जीनोमी डेटाबेसों को एकीकृत करने के लिए जेनेरिक बनाया गया है। यह पोर्टल अब उपयोगकर्ताओं द्वारा उनके जीव आंकड़े प्रस्तुत किए जाने के लिए खुला हुआ है। आंकड़ों की गुणवत्ता के प्रबंध के लिए

एक ऑटो–क्यूरेशन कार्यक्रम भी विकसित किया जा रहा है। आवश्यकता है कि हमारे अनुसंधानकर्ताओं को इस पोर्टल के बारे में विश्वास सृजित करते हुए उपयोग के लिए प्रोत्साहित किया जाए जिसके लिए उनके डेटा को सुरक्षा प्रदान करने और प्रौद्योगिकी एवं कृषि उत्पादकता में सुधार के लिए ज्ञान प्राप्त करने हेतु इन आंकड़ों की भागीदारी करनी होगी।

बकरी प्रजनन पहचान सर्वर : माइक्रोसेटेलाइट डीएनए मार्कर का उपयोग करके एक बकरे–बकरियों की नस्ल की पहचान के लिए वैब आधारित सर्वर विकसित किया गया है। यह बायेसियन नेटवर्क्स क्लासीफायर पर आधारित है जिसकी सटीकता 98.7 प्रतिशत है और इसमें भारत की बकरे–बकरियों की 22 नस्लों पर 25 माइक्रोसेटेलाइट स्थलों द्वारा 51850 संदर्भ युग्मविकल्पी आंकड़े सृजित किए गए हैं– (http://webapp.cabgri.res.in/gomi/) (इकबाल एवं साथी, 2014)। सामान्यतः नस्ल संबंधी विवरण केवल 'शुद्ध नस्ल' प्रकार के पशुओं के लक्षण–वर्णन के लिए किए गए हैं जिसमें गैर–पहचाने गए या मिश्रित नस्ल की जनसंख्या को शामिल नहीं किया गया है। इसके अतिरिक्त वीर्य, डिम्ब, भ्रूण तथा नस्ल उत्पाद के मामले में दृष्टव्य विवरणों की कमी के कारण नस्ल की पहचान नहीं की जा सकती है। इसलिए माइक्रोसेटेलाइट तथा एस.एन.पी. जैसे आण्विक मार्करों के आने से इनका उपयोग छोटे जीवविज्ञान ऊतक या जननद्रव्य से भी नस्ल की पहचान के लिए किया जा सकता है। इस सर्वर से लागत तो कम होगी ही, कम्प्यूटेशनल कामों में भी आसानी होगी। यह क्रियाविधि समस्त फ्लोरा और फाना के लिए एक उदाहरण बन जाएगी जो सीमापार जननद्रव्य को लाने–ले जाने व प्रबंध में सार्वभौमिकता संबंधी मुद्दों को हल करने के साथ–साथ संरक्षण, नस्ल सुधार कार्यक्रमों के लिए एक उपयोगी तकनीक सिद्ध होगी।

गोपशु प्रजनन पहचान सर्वरः गुणप्ररूपी विवरणों विशेष रूप से डिम्ब, वीर्य, भ्रूण और प्रजनन उत्पादों की कमी, अपमिश्रण के अंश के उचित रूप से न पता लगा पाने व गैर वर्णित पशुओं के कारण, नस्लों की पहचान के लिए संदर्भ आण्विक आंकड़े न होने आदि के कारण गोपशु नस्ल की पहचान से जुड़ी चुनौतियों से निपटने के लिए यह सर्वर विकसित किया गया है। साथ ही यह वैब सर्वर संदर्भ आंकड़ों के रखरखाव तथा गोपशु नस्ल की पहचान के लिए विकसित किया गया है– (http://webapp.cabgrid.res.in/biscattle/)। इन संदर्भ आंकड़ों का उपयोग 8 गोपशु नस्लों तथा 18 माइक्रोसेटेलाइट डीएनए मार्करों से प्राप्त पूर्वानुमान मॉडल के विकास के लिए किया गया है जिससे 18000 युग्मविकल्पी आंकड़े प्राप्त हुए हैं। महत्वपूर्ण स्थलों की पहचान के लिए या स्थलों की संख्या कम करने के लिए विभिन्न एल्गोरिथ्म का उपयोग किया गया। मैमोरी–आधारित अधिगम एल्गोरिथ्म का उपयोग करके स्थलों की संख्या को 5 तक कम किया गया और इसमें 95 प्रतिशत सटीकता भी बनी रही। यह मॉडल दृष्टिकोण, क्रियाविधि नस्ल की पहचान तथा संरक्षण कार्यक्रम में पूरे विश्व के सभी पालतू पशु प्रजातियों की पहचान करने में महत्वपूर्ण भूमिका निभा सकता है।

वर्क फ्लो पाइपलाइन तथा जीवविज्ञानी आंकड़ा विश्लेषण के लिए युक्तियां : प्रतिसूक्ष्मजैविक पैप्टाइड (एएमपी) प्रतिरक्षा अणु हैं तथा रासायनिक प्रतिजैविकों के प्राकृतिक विकल्प हैं। मशीन द्वारा सीखने की तकनीकें बृहत जीवविज्ञानी आंकड़ों में छिपे हुए पैटर्नों को समझने के लिए पर्याप्त उपयोगी सिद्ध हुए हैं। यह पाया गया है कि गोपशुओं के एएमपी *के इन–सिलिको* पूर्वानुमान/पहचान के लिए एसवीएम आधारित मॉडल, एएनएन की तुलना में निष्पादन के मामले में श्रेष्ठ हैं। विभिन्न डेटाबेसों से एकत्र किए गए गोपशुओं से संबंधित

कुल 99 एएमपी को साहित्य के रूप में प्रकाशित किया गया है। एसवीएम मॉडल विकास तथा पहचान/पूर्वानुमान के लिए एन–टर्मिनस रेसिड्यू, सी–टर्मिनस रेसिड्यू तथा पूर्ण क्रमों का उपयोग किया गया है (सारिका और साथी, 2015)। ये एसवीएम मॉडल वैब सर्वर पर कार्यान्वित किए गए तथा गोपशुओं के नए एएमपी के वर्गीकरण/पूर्वानुमान के लिए http://webapp.cabgri.res.in/amp/ पर उपयोगकर्ताओं को उपलब्ध कराए गए हैं।

वर्कफ्लो तथा पाईपलाइन जीन पूर्वानुमान, फाइलोजेनेटिक विश्लेषण तथा एसएसआर–प्राइमर के लिए एक समानांतर फ्रेमवर्क विकसित किया गया है जो सार्वजनिक क्षेत्र में उपलब्ध विभिन्न युक्तियों के एकीकरण के माध्यम से संभव हुआ है। जीन अभिव्यक्ति की पहचान हेतु कोडोन उपयोग के विश्लेषण के लिए एक वैब आधारित सॉफ्टवेयर कार्यान्वित किया गया है। जीनों से संबंधित विशेषकों (उदाहरणतः अजैविक या जैविक प्रतिबल) का पूर्वानुमान जीवविज्ञानी अनुसंधान के प्रयोगकर्ताओं के लिए बहुत उपयोगी है। जीन अभिव्यक्ति संबंधी आंकड़े उपयोगी हो सकते हैं लेकिन इसके लिए विशेषज्ञतापूर्ण विश्लेषणात्मक तथा कम्प्यूटेशनल सहायता की आवश्यकता होती है। गुण विशेषक संबंधित जीन पूर्वानुमान युक्ति (टीएजीपीटी) जो एक उपयोगकर्ता मित्र वैब आधारित विश्लेषण समाधान है तथा वैज्ञानिकों द्वारा विकसित की गई है। टीएजीपीटी ठोस सांख्यिकी सिद्धांतों पर आधारित प्रस्तावित एल्गोरिथ्म को कार्यान्वित करता है तथा इसके लिए इनपुट के रूप में जीन अभिव्यक्ति संबंधी आंकड़ों की आवश्यकता होती है। वर्तमान में वैज्ञानिक जीन विनियमनकारी नेटवर्क (जीआरएन) की मॉडलिंग के लिए वैब आधारित युक्ति विकसित करने का कार्य कर रहे हैं। इस ऑन लाइन तकनीक से पूर्व संसाधित अगली पीढ़ी के अनुक्रमण (एनजीएस)/माइक्रोऐरे आंकड़े प्राप्त करने, विभिन्न मॉडलिंग फार्मेलिज़्म द्वारा जीआरएन का निर्माण करने तथा नेटवर्क को देखने की सुविधा प्राप्त होगी। यह कार्यक्रम माइक्रोऐरे तथा एनजीएस आंकड़ों से जीन अभिव्यक्ति की गणन करता है जिसका उपयोग विनियमनकारी नेटवर्कों की पुनर्संरचना करने और उसके द्वारा इन्हें देखने में किया जा सकता है। प्रोटीन संरचना की तुलना (पीएससी) प्रोटीनों के बीच के विकासात्मक संबंधां को समझने, प्रोटीनों की संरचना व कार्य के पूर्वानुमान को भी समझने के लिए एक महत्वपूर्ण कार्य है। समांगी प्रोटीनों का पता लगाने, उनके कार्यात्मक वर्गीकरण व संरचनात्मक मॉटिफों की खोज के लिए इन प्रोटीनों की संरचनाओं की तुलना की जाती है। प्रोटीन संरचनाओं की तुलना के लिए अनेक विधियां प्रस्तावित की गई हैं और प्रत्येक विधि में उस विधि की स्कोरिंग स्कीम को उपयुक्ततम किया जाता है। एक कुशल एल्गोरिथ्म पर आधारित वैब आधारित तकनीक उपयोगकर्ताओं के लिए विकसित की गई है जिससे प्रोटीन संरचना की तुलना की जा सकती है। प्रोटीन 3डी संरचनाओं की मात्रात्मक तुलना संरचनात्मक जीवविज्ञान में एक महत्वपूर्ण तथा मूलभूत कार्य है जिससे अन्य प्रोटीनों के साथ विकासात्मक तथा संरचनात्मक संबंधों का अध्ययन किया जाता है। इससे जीवविज्ञानियों को संरचनात्मक पड़ोसी प्रोटीनों के कार्य, विकास संबंधी विभिन्न पहलुओं को समझने व संरचनाओं की पहचान करने में सहायता मिलती है। अब त्रिआयामी प्रोटीन संरचनाओं का डेटाबेस बड़ा होता जा रहा है, अतः तेजी से और सटीक रूप से खोज करने वाली तकनीकों तथा तुलना करने वाली विधियों की अत्यंत आवश्यकता है। संरचना की तुलना संरचनाओं के विश्लेषण व उनकी खोज द्वारा संरचना की विविधता को समझने में मुख्य भूमिका निभा सकती है ताकि रूचिकर वैज्ञानिक अंतर दृष्टियां सृजित हो सकें। साहित्यिक सर्वेक्षण के आधार पर 3डी प्रोटीन संरचना के मात्रात्मक निर्धारण तथा उनकी युग्मवार तुलना के लिए ग्राफ सैद्धांतिक तकनीक का उपयोग किया जा सकता है। वैज्ञानिक

3 डी प्रोटीन संरचना के मात्रात्मक निर्धारण व ग्राफ सैद्धांतिक तकनीक का उपयोग करके युग्मवार तुलना के लिए कारगर एल्गोरिथ्म विकसित करने के कार्य में लगे हुए हैं।

ऐसी आशा है कि शीघ्र ही जैव सूचना के क्षेत्र में अनुसंधान करने के लिए पर्याप्त प्रशिक्षित जनशक्ति उपलब्ध होगी जिससे कृषि जैवप्रौद्योगिकी में अनुसंधान व विकास संबंधी सहायता प्राप्त हो सकेगी। इन जीनोमी आंकड़ों से प्राप्त ज्ञान को उपलब्ध कराने के लिए उच्च स्तर की कम्प्यूटेशनल क्षमता से युक्त आंकड़ा खोजने की सुविधाओं व केन्द्रीय जीनोमी डेटा वेयरहाउसिंग (सीजीडीडब्ल्यू) का हमारे देश में विकास होगा। हम इस राष्ट्रीय ग्रिड में एनएआरईएस की अन्य संस्थाओं को भी शामिल करने के कार्य में लगे हुए हैं जो प्रावस्थावार होगा तथा कृषि अनुसंधान की प्राथमिकता पर निर्भर करेगा। इससे हम न केवल अपने देश से बल्कि पूरे विश्व से भूख को दूर कर सकेंगे।

संदर्भ

1. इकबाल एम.ए., सारिका, अरोड़ा वसु, वर्मा निधि, राय अनिल और कुमार दिनेश (2013). फर्स्ट होल जीनोम बेस्ड माइक्रोसेटेलाइट डीएनए मार्कर डेटाबेस ऑफ टोमेटो फार मैपिंग एंड वेराइटी आइडेंटिफिकेशन. बीएमसी प्लांट बायोलॉजी 2013, 13रु197. डीओआई 10. 1186 / 1471–2229–13–107.

2. इकबाल एम.ए., अंसारी, एम.एस., सारिका, दीक्षित, एस.पी., वर्मा, एन.के., अग्रवाल, आर. , ए.के., जयकुमार, एस., राय, ए. और कुमार डी, (2014). लोकस मिनिमाइजेशन इन ब्रीड प्रीडक्शन यूजिंग आर्टिफिसियल न्यूट्रल नेटवर्क एप्रोच. एनिमल जेनेटिक्स, 45(6), 898–902, नास स्कोरः 8.21.

3. लाल, एस.बी., पांडे, पी.के., राय, पी.के., राय, ए., शर्मा, ए, चतुर्वेदी, के.के. (2013). डिजाइन एंड डेवलपमेंट ऑफ पोर्टल फॉर बायोलॉजिकल डेटाबेस इन एग्रीकल्चर. बायोइंफोर्मेशन, 9(11): 588–598.

4. नवीन शर्मा, मोहम्मद समीर फारूकी, कृष्ण कुमार चतुर्वेदी, शशि भूषण लाल, मोनेन्द्रा ग्रोवर, अनिल राय और पंकज पाण्डे (2014). द हैलोफाइल प्रोटीन डेटाबेस. डेटाबेस (ऑक्सफोर्ड) : द जर्नल ऑफ बायोलॉजिकल डेटाबेसिस एंड क्यूरेशन, डीओआई : 10. 1093 / डेटाबेस / बी.एयू .114.

5. सारिका, अरोड़ा, वसु, इकबाल, एम.ए., राय, अनिल और कुमार, दिनेश (2013). इन सिलिको माइनिंग ऑफ प्यूटेटिव माइक्रोसेटेलाइट मार्कर्स फ्राम होल जीनोम सीक्वेंस ऑफ वाटर बफलो (*बैबुलस बैबुलिस*) एंड डेवलपमेंट ऑफ फर्स्ट बफसैट डीबी.बीएमसी जीनोमिक्स. 14, 43, डीओआई : 10.1186 / 1471–2164–14–43.

6. सारिका, अरोड़ा, वसु, इकबाल, एम.ए., राय, अनिल और कुमार, दिनेश (2013). पीआईपीई माइक्रो डीबी : माइक्रोसेटेलाइट डेटाबेस एंड प्राइमर जनरेशन टूल फार पीजनपी जीनोम. डेटाबेस : द जर्नल आफ बायोलॉजिकल डेटाबेसिस एंड क्यूरेशन. खंड 2013, लेख आइडी बीएएस054, डीओआईः 10.1093 / डेटाबेस / बीएएस054.

7. सारिका, इकबाल, एम.ए., अरोड़ा, वसु, राय, अनिल और कुमार, दिनेश (2015). स्पीसिज स्पेसिफिक एप्रोच फॉर डेवलपमेंट ऑफ वैब बेस्ड एंटिमाइक्रोबायल पेप्टाइड्स प्रेडिक्शन टूल. कम्प्यूटर एंड इलेक्ट्रोनिक्स इन एग्रीकल्चर, 111, 55–61, नास स्कोर 7.49.

अशोका - एक परिचय

जैव सूचना विज्ञान जीव विज्ञान, संगणक विज्ञान, गणित विज्ञान एवं सांख्यिकी विषयों के आपसी सहयोग से मिलकर बना है। जैव सूचना विज्ञान जीन एवं उनके कारकों के कार्य को समझने, कृषि उत्पादकता बढ़ाने, उन्नत किस्मों एवं नस्लों के विकास में सहायक है। जेनेटिक इंजीनियरिंग और जीनोमिक दृष्टिकोण से कृषि सम्बन्धी उत्पादों की उत्पादकता और गुणवत्ता की विशेषताओं को बढ़ाने के लिए, जैव सूचना विज्ञान का एक नए विषय के रूप में सृजन हुआ हैं।

दुनिया भर में आणविक प्रयोगशालाओं के विकास एवं जैविक अनुक्रमण प्रौद्योगिकियों में प्रगति के कारण, बहुत अधिक मात्रा में जैविक आंकड़े उत्पन्न हो रहे हैं। सूचना और संचार प्रौद्योगिकी के क्षेत्र में विकसित नई तकनीकियां इन आंकड़ों को एकत्रित, संग्रहित, संचित एवं विश्लेषित करने में सहायक सिद्ध हो सकती हैं। मूर के नियमानुसार, कंप्यूटर की गणना करने की क्षमता डेढ़ से दो महीनों में दुगुनी हो जाती है। जैविक आंकड़ों में छिपे हुए जैविक ज्ञान को निकालने के लिए उच्च प्रदर्शन कंप्यूटिंग सुविधाओं की जरूरत है। उच्च प्रदर्शन कंप्यूटिंग या हाई परफॉरमेंस कंप्यूटिंग (एचपीसी) आंकड़ों को जल्दी, कुशलतापूर्वक एवं उन्नत एप्लीकेशन सॉफ्टवेयर की सहायता से विश्लेषित कर सकता है। एचपीसी की क्षमता का आंकलन फ्लॉप्स (FLOPS - Floating point operations per second) में किया जाता है। एचपीसी तकनीकी रूप से एक सुपर कंप्यूटर के रूप में सबसे ज्यादा प्रचलित हुआ है।

भारतीय कृषि अनुसंधान परिषद (आई.सी.ए.आर.) ने भा.कृ.अनु.प.–भारतीय कृषि सांख्यिकी अनुसंधान संस्थान, नई दिल्ली में कृषि जैव सूचना विज्ञान केंद्र की स्थापना की है। भारतीय कृषि अनुसंधान परिषद (आई.सी.ए.आर.) ने विश्व बैंक द्वारा पोषित राष्ट्रीय कृषि नवोन्मेषी परियोजना (एन.ए.आई.पी.) के अंतर्गत भा.कृ.अनु.प.–भारतीय कृषि सांख्यिकी अनुसंधान संस्थान, नई दिल्ली में एक उप परियोजना राष्ट्रीय कृषि जैव सूचना ग्रिड (एन.ए.बी.जी.) की आईसीएआर में स्थापना की स्वीकृत की। इस परियोजना में भा.कृ.अनु.प.–राष्ट्रीय पादप आनुवंशिक संसाधन ब्यूरो (एन.बी.पी.जी.आर.) नई दिल्ली, भा.कृ.अनु.प.–राष्ट्रीय पशु आनुवंशिक संसाधन ब्यूरो (एन.बी.एजी.आर.) करनाल, भा.कृ.अनु.प.–राष्ट्रीय मत्स्य आनुवंशिक संसाधन ब्यूरो (एन.बी.एफ.जी.आर.) लखनऊ, भा.कृ.अनु.प.–राष्ट्रीय कृषि उपयोगी सूक्ष्मजीवों ब्यूरो (एन. बी.ऐ.आई.एम.) मऊ और भा.कृ.अनु.प.–राष्ट्रीय कृषि कीट संसाधन ब्यूरो (एन.बी.ऐ.आई.आर.), बंगलुरु सहयोगी संस्थान थे। इस परियोजना का प्रमुख उद्देश्य जैव आंकड़ों के विश्लेषण हेतु उच्च प्रदर्शन कंप्यूटिंग या हाई परफॉरमेंस कंप्यूटिंग (एचपीसी) की स्थापना एवं जीनोमिक डेटा संसाधनों और विभिन्न जैविक डेटाबेस के विकास, विश्लेषण और भंडारण करना है ।

संस्थान में एचपीसी की स्थापना मिश्रित रूप में की गयी है । इसमें चार अलग अलग अर्थात् 40 नोड्स लाइनक्स, 16 नोड्स जी पी–जी पी यू लाइनक्स, 16 नोड्स बिग डाटा, एक 1.5 टीबी रैम और एक 1.0 टीबी रैम से निहित सममित बहु–प्रोसेसर (एसएमपी) के रूप में सुपरकंप्यूटर स्थापित किये गए हैं । आंकड़ों को रखने एवं विश्लेषित करने हेतु भंडारण क्षमता को तीन घटकों (अ) नेटवर्क फाइल सिस्टम (ब) समानांतर फाइल सिस्टम और (स) संग्रह प्रणाली (आर्काइवल) में विभाजित किया गया है। इन सभी को जोड़ने के

लिए तीन प्रकार के नेटवर्क बनाये गए हैं। क्यू–लॉजिक का उच्च बैंडविड्थ नेटवर्क (क्यूडीआर इनफिनीबैंड स्विच) सभी नोड्स एवं भण्डारण क्षमता प्रणाली के बीच में सम्बन्ध स्थापित कर एक दूसरे को सन्देश पहुँचाने में सहायता करता है। गीगाबिट नेटवर्क का उपयोग क्लस्टर प्रशासन और प्रबंधन के लिए किया गया है। आईएलओ–3 नेटवर्क का उपयोग समस्त नोड्स अन्य उपकरणों के स्वास्थ्य की निगरानी एवं प्रबंधन के लिए किया गया है। प्रत्येक सहयोगी संस्थान में भी एक 16 नोड्स लाइनक्स सुपरकंप्यूटर स्थापित किया है । इन पांचों संस्थानों के सुपरकंप्यूटर को भी मुख्य संस्थान के साथ एम पी एल एस कनेक्टिविटी के द्वारा एकीकृत किया गया है (चित्र १)। ये सुपर कंप्यूटर उपयोगकर्ताओं के लिए एक मिश्रित वास्तुकला का अनूठा उदाहरण प्रस्तुत करते हैं ।

इस सुपरकंप्यूटर का नाम अशोका (ASHOKA: Advanced Supercomputing Hub for Omics Knowledge in Agriculture) दिया गया है। यह सुविधा जैव सूचना उपकरण, डेटाबेस निर्माण और उनके उपयोग से जैविक अनुसंधान को उन्नत बनाने में सहायक सिद्ध हो रहा है । अशोका को कमांड लाइन इंटरफेस (सी.एल.आई.) और वेब पोर्टल की सहायता से प्रयोग कर सकते हैं। अशोका प्रयोग करने लिए सर्वप्रथम पंजीकरण करना अनिवार्य है। पंजीकरण बायो–कंप्यूटिंग पोर्टल के माध्यम से किया जा सकता है। बायो–कंप्यूटिंग पोर्टल चित्र 1 में दर्शाया गया है।

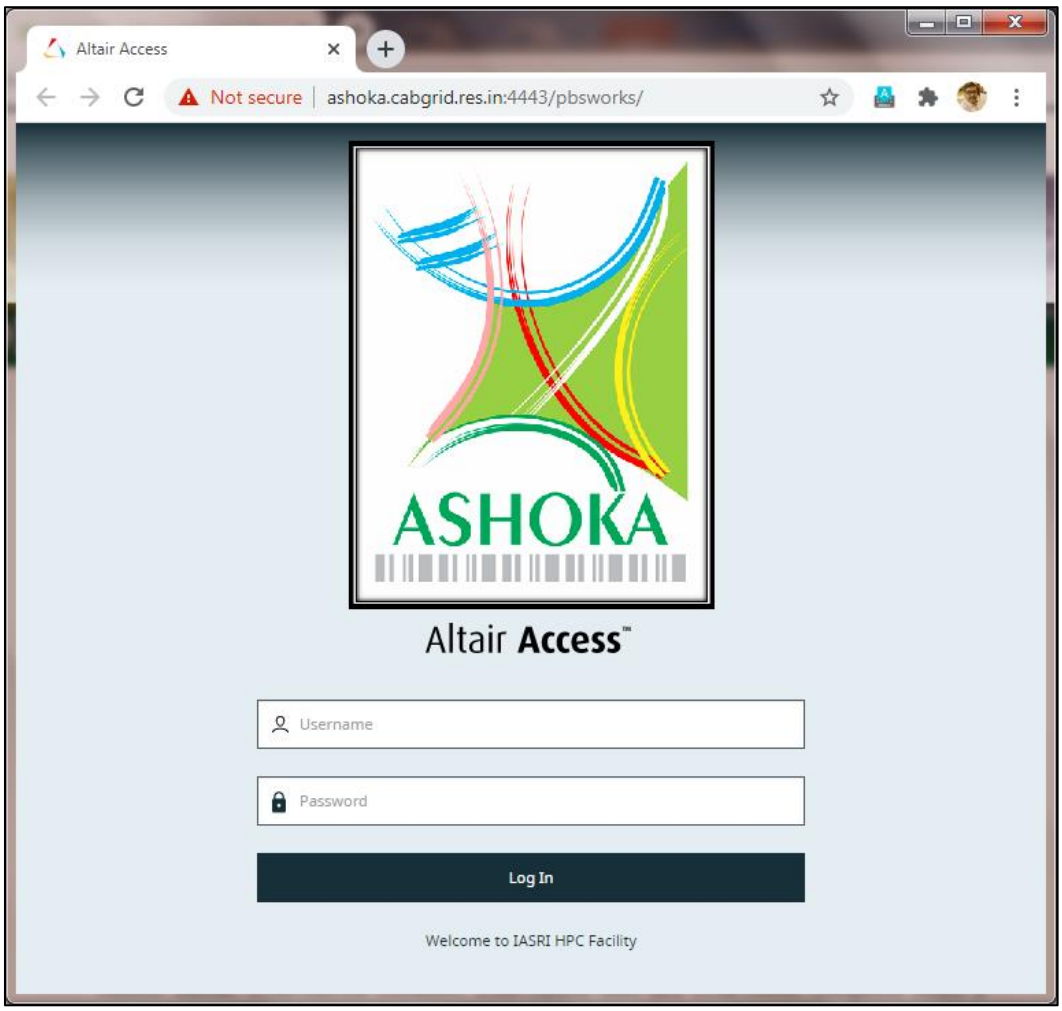
चित्र 1 : बायो–कंप्यूटिंग पोर्टल

प्रयोगकर्ता को उनकी आवश्यकतानुसार वर्गीकृत किया है जो कि निम्न प्रकार से हैं

1. केंद्र उपयोगकर्ता – प्रभाग में कार्यरत वैज्ञानिकों एवं शोध सहायकों के लिए
2. परिषद् उपयोगकर्ता – परिषद् के संस्थानों में कार्यरत वैज्ञानिकों एवं शोध सहायकों के लिए

3. अन्य उपयोगकर्ता (आर यू) – अन्य संस्थान के वैज्ञानिकों एवं शोध सहायकों के लिए

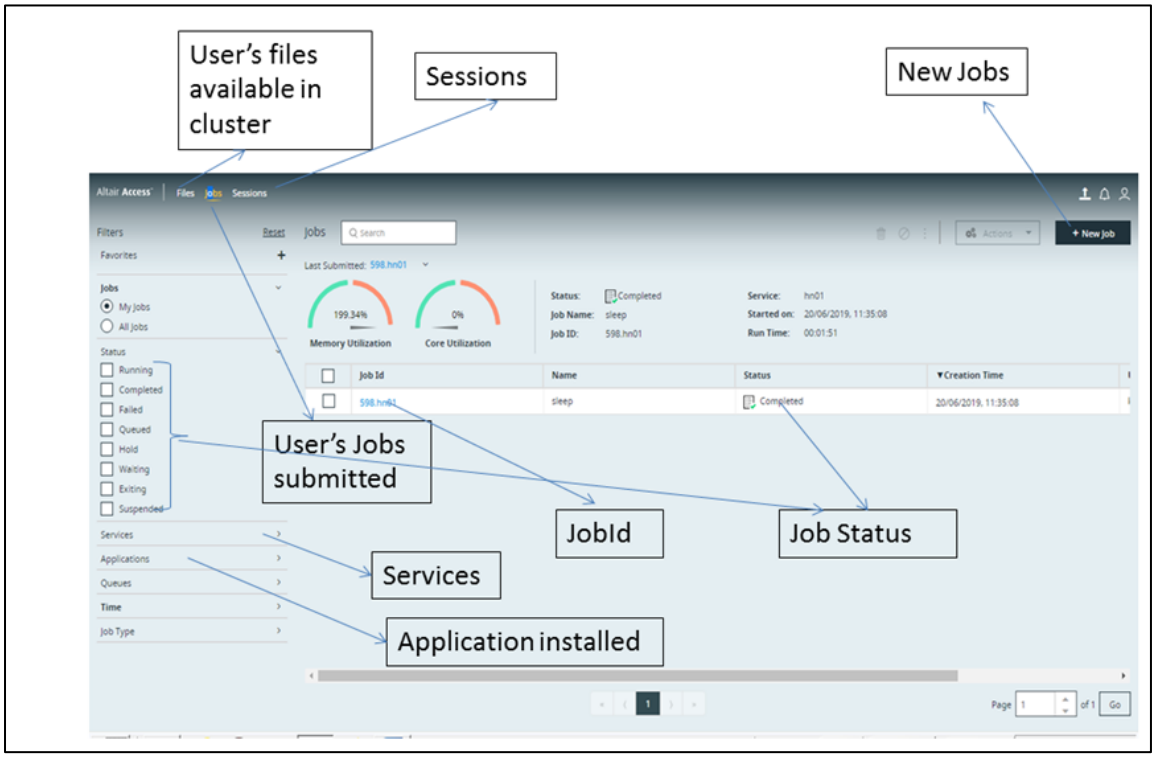
उपयोगकर्ताओं की अनुसंधान गतिविधियों और आवश्यकताओं के अनुसार उनके वर्ग को बदला भी जाता है। इस पोर्टल में विभिन्न मुक्त स्रोत सॉफ्टवेयर का मैत्रीपूर्ण ग्राफिकल यूजर इंटरफेस (जी यू आई) बनाया गया है जोकि जैविक वैज्ञानिकों एवं शोध सहायकों के लिए अत्यंत उपयोगी है । अशोका में लोगिन और प्रवेश करने के लिए वेबपेज को चित्र २ में दर्शाया गया है। पोर्टल प्रमाणीकृत उपयोगकर्ताओं को अपने–अपने स्थानों से उनके जैविक डेटा विश्लेषण एवं प्रदर्शन करने में सक्षम है। इसके विकास के दौरान उपयोगकर्ता की आवश्यकताओं को ध्यान में रख कर निर्मित किया गया है।



चित्र 2 : अशोका में लोगिन और प्रवेश

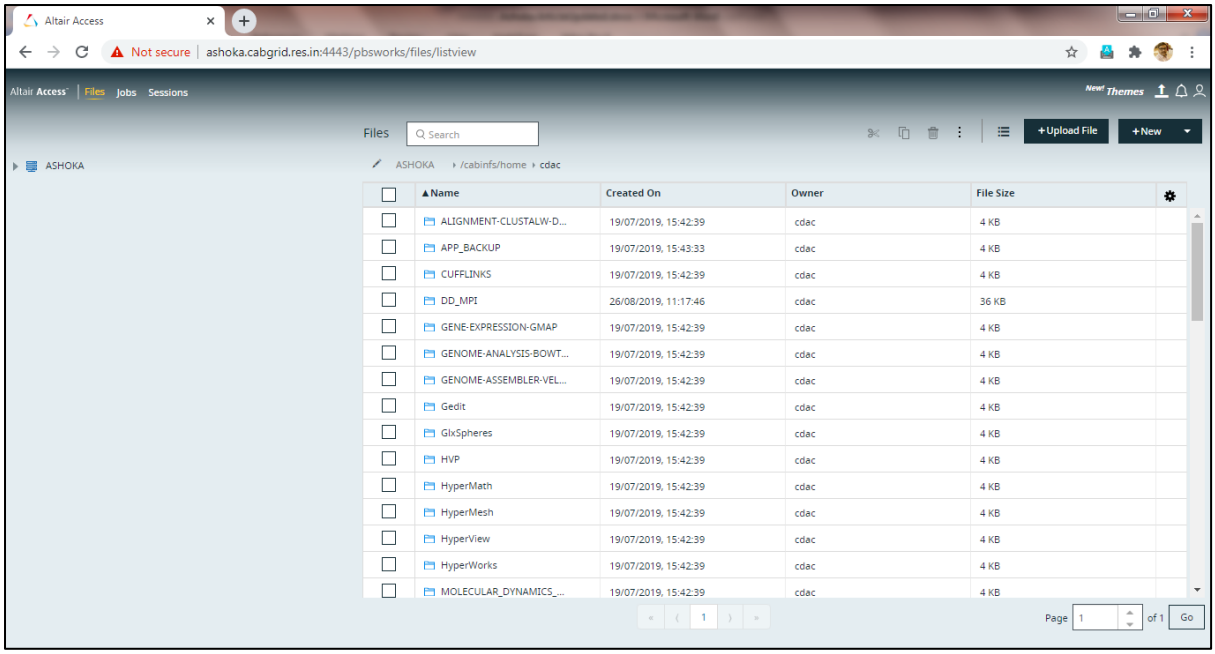
उपयोगकर्ता सफलतापूर्वक लोगिन करने के पश्चात, वह अपनी जॉब प्रस्तुत, निगरानी और प्रबंध कर देख सकता है और जॉब की प्रगति को देखते हुए उचित निर्णय भी ले सकता है कि जॉब को आगे चलना चाहिए या बंद कर देना चाहिए। कई महत्त्वपूर्ण सॉफ्टवेयर को उपयोगकर्ताओं की आवश्यकताओं को ध्यान में रखते हुए अक्सर इस्तेमाल में आने वाले विकल्पों के साथ पोर्टल में एकीकृत किया गया है। आणविक और आनुवंशिक प्रक्रिया से संबंधित कृषि अनुसंधान उपलब्ध आंकड़ों एवं परिणामों को सांख्यिकीय और कम्प्यूटेशनल तकनीकों की सहायता से विश्लेषित करने में सहायक है।

जॉब प्रस्तुतीकरण (जॉब सबमिशन): प्रस्तुत मॉड्यूल जॉब को संगठित करने और पंजीकृत सर्वरों में प्रस्तुत करने के लिए प्रयोग किया जाता है। इस तरह के मॉड्यूल में एक जॉब के रूप में कुछ सामान्य विकल्प शामिल हैं और इनपुट ब्राउज करने की क्षमता को स्थानीय रूप अच्छी तरह से सहारा देता है और उपयोगकर्ताओं के लिए लाभकारी है। उपयोगकर्ता आवश्यक जॉब प्रपत्र का चयन कर जॉब निष्पादन के लिए इस्तेमाल करने की अनुमति देता है (चित्र 3) । आवेदन से संबंधित चयन करने पर सभी अनिवार्य और वैकल्पिक विकल्प के रूप में अच्छी तरह से प्रदर्शित हो रहे हैं। फाइल प्रबंधन सेवाओं के उपयोगकर्ताओं, इनपुट फाइल /फोल्डर और लिपियों डाउनलोड, आवेदन उत्पादन फाइलें / फोल्डर को अपलोड फाइलों और फोल्डरों को हटाना, एक दूसरे के लिए एचपीसी संसाधन से नकल करने की अनुमति देता है (चित्र 4) ।



चित्र 3 : जॉब निगरानी और प्रबंध

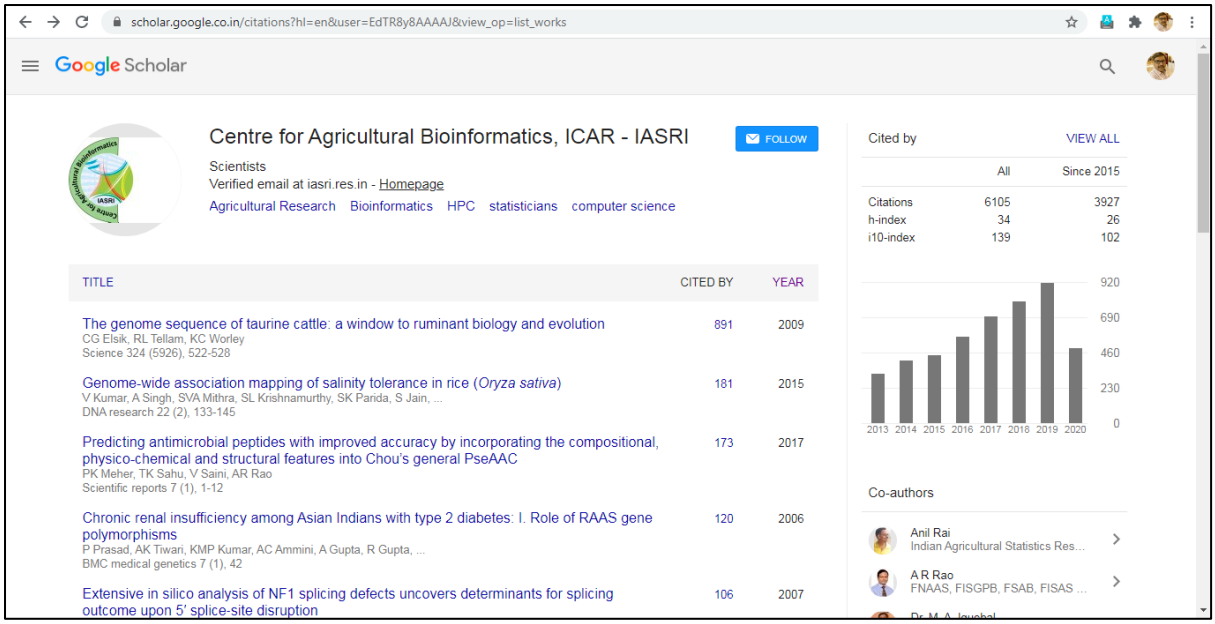
जॉब निरीक्षण (जॉब मॉनीटरिंग): उपयोगकर्ता जॉब प्रस्तुत के बाद जॉब की निगरानी एवं प्रबंध भी कर सकता है । पोर्टल प्रस्तुत जॉब की वर्तमान स्थिति के बारे में जानकारी प्रदान करता है और कोई एक निर्णय लेने के विकल्प भी प्रदान करता है। उपयोगकर्ता आसानी से विभिन्न स्थितियों के साथ इन जॉब्स का प्रबंध कर सकते हैं। जॉब प्रबंध करने के लिए जॉब आईडी, जॉब का नाम, जॉब का मालिक, जॉब की वर्तमान स्थिति, कतार, आवंटित संसाधन, वाल समय, निष्पादन नोड्स इत्यादि को पोर्टल के द्वारा नियंत्रित कर सकता है।



चित्र 4: फाइल प्रबंधन

जॉब विश्लेषण (जॉब एनालिटिक्स): पोर्टल का संरक्षक (एडमिनिस्ट्रेटर) किसी भी जॉब का स्टेटस पता लगा सकता है तथा एच पी सी में चल रही जॉब्स की प्रगति के बारे में भी जान सकता है। जॉब विश्लेषण मॉड्यूल सभी जॉब्स को एकीकृत रूप से विभिन्न ग्राफ एवं चार्ट्स के माध्यम से समझने में सहायता प्रदान करता है (चित्र ७)। इस पोर्टल की सहायता से विभिन्न पहलुओं पर एकीकृत जानकारी भी प्राप्त की जा सकती है जैसे कि (अ) कितनी जॉब्स चल रही हैं, (ब) कितनी जॉब्स समाप्त गयी हैं, (स) कितनी जॉब्स कतार में हैं, (द) प्रतिदिन कितनी जॉब्स चलती हैं (इ) कितने नोड्स फ्री और व्यस्त हैं और कई अन्य।

अशोका सुपर–कम्प्यूटर के माध्यम से कृषि जैव सूचना केंद्र ने संसथान एवं परिषद् में नए आयाम स्थापित किये हैं तथा कई उच्च कोटि के शोध पत्र भी प्रकाशित किये हैं । उच्च कोटि के प्रकाशनों की उपयोगिता को गूगल स्कॉलर के माध्यम से देखा जा सकता है (चित्र 5)।



चित्र 5 : प्रकाशनों की उद्धरण

(स्रोतः गूगल स्कॉलर
https://scholar.google.co.in/citations?hl=en&user=EdTR8y8AAAAJ&view_op=list_works)

जैव प्रौद्योगिकी बैक्टीरिया, वायरस, कवक, आदि खमीर, पशु कोशिकाओं, संयंत्र कोशिकाओं को बनाने या पौधों या जानवरों में सुधार करने के लिए या विशिष्ट उपयोगों के लिए सूक्ष्म जीवों को इंजीनियर करने के लिए अनिवार्य विषय हो गया है । विगत वर्षों में उच्च प्रदर्शन कंप्यूटिंग (एच पी सी) ने बृहद आंकड़ों को विश्लेषित करने में एक महत्तवपूर्ण योगदान दिया है और अशोका का निर्माण अपने देश के जैव वैज्ञानिकों के लिए सहायक एवं लाभकारी सिद्ध होगा । इस बड़े पैमाने पर जैविक डेटा में एन्क्रिप्टेड जैविक ज्ञान को समझने के लिए उच्च प्रदर्शन कम्प्यूटेशनल बुनियादी ढांचे में डेटा एकीकरण, फ्यूजन, खनन, कार्यप्रवाह विकास और निष्पादन, उद्गम और प्रतिनिधित्व करने के लिए इस्तेमाल किया जा सकता है। अशोका जीनोमिक डेटा संसाधनों और विभिन्न जैविक डेटाबेस के दृश्य का विश्लेषण और भंडारण के लिए उपयोगी है।

बेसिक लोकल एलाइनमेंट सर्च टूल

प्रस्तावना

मौलिक स्थानीय एलाइनमेंट खोज युक्ति अथवा बेसिक लोकल एलाइनमेंट सर्च टूल (ब्लास्ट) सभी प्रमुख सीक्वेंस डेटाबेसों के खोजने की एक लोकप्रिय उपयोगकर्ता मित्र युक्ति है। ब्लास्ट क्वेरी सीक्वेंस तथा डेटाबेस सीक्वेंस के बीच स्थानीय रूप से उपयुक्ततम एलाइनमेंट में सर्वाधिक स्कोरिंग का पता लगाने की अनुमानी विधि है। ब्लास्ट कार्यक्रमों को त्वरित डेटाबेस खोजने के लिए डिजाइन किया गया था जिसके अंतर्गत सुदूर संबंधित क्रमों की संवेदनशीलता का कम से कम त्याग करना पड़ता था। ब्लास्ट का उपयोग क्वेरी सीक्वेंस को पहचानने, उसके कार्यों तथा 3डी संरचना के पूर्वानुमान के लिए सीक्वेंस समांगों का पता लगाने के लिए किया जाता है। ब्लास्ट से न्यूक्लियोटाइड सीक्वेंसों की तुलना में प्रोटीन सीक्वेंसों के बेहतर परिणाम प्रदर्शित होते हैं।

लोकल एलाइनमेंटों की आवश्यकता

लोकल एलाइनमेंट उच्च यूकैरियोटों से नए डी.एन.ए. कामों के मामले में विशेष रूप से महत्वपूर्ण हैं। सर्वाधिक उल्लेखनीय एलाइनमेंट प्रोटीन डेटाबेसों के विरुद्ध नए डी.एन.ए. सीक्वेंस के ट्रांसलेशन हैं जिनका लाभ बड़े प्रोटीन एल्फाबेट में उठाया जा सकता है तथा इनसे न्यूक्लिओटाइड संरेखणों (Nucleotied Allignments) के 25 प्रतिशत बेतरतीब एलाइनमेंट गुणों से बचा जा सकता है। तथापि किसी नए डी.एन.ए. सीक्वेंस में जीन एक्सॉन के बीच वितरित होते हैं और किसी लंबे डी.एन.ए. सीक्वेंस में एक से अधिक जीन हो सकते हैं। इसके अतिरिक्त कुछ इन्ट्रॉन में स्वयं उनके जीन होते हैं जैसे न्यूक्लिऐज; इस प्रकार नए डी.एन.ए. में 'जीनों के अंदर जीन' हो सकते हैं।

इस प्रत्येक विशेषता के लिए स्थानीय एलाइनमेंटों की आवश्कयता होती है जिससे प्रत्येक जीन के लिए एलाइनमेंट प्राप्त होगा और इससे भी अधिक महत्वपूर्ण यह है कि प्रत्येक एक्सॉन के लिए भी एलाइनमेंट प्राप्त हो सकता है।

ब्लास्ट कार्यक्रमों से इन्हें हल किया जा सकता है।

स्थानीय समानता - किसी अंतराल की अनुमति नहीं।

गणितीय रूप से कठिनः स्कोरों का वितरण एक अत्यधिक मूल्यवान वितरण है।

एल्गोरिथ्म में लुकअप तालिकाओं का उपयोग होता है तथा छोटे 'शब्द' का उपयोग उनसे मेल खाता है।

ब्लास्ट की कुछ प्रमुख विशेषताएं हैं :

स्थानीय एलाइनमेंटः ब्लास्ट क्वेरी तथा डेटाबेस सीक्वेंस के बीच वैश्विक फिट होने की बजाय स्थानीय समानता के पैचों का पता लगाने का प्रयास करता है।

अंतरालहीन एलाइनमेंटः ब्लास्ट कार्यक्रम अंतरालहीन सीक्वेंस एलाइनमेंट्स की सांख्यिकी पर कार्य करते हैं लेकिन सैद्धांतिक रूप से इससे खोज की संवेदनशीलता कम हो जाती है। तथापि, आउटपुट में क्वेरी तथा डेटाबेस सीक्वेंस के बीच अनेक स्थानीय एलाइनमेंट प्रदर्शित होते हैं जिनका उपयोग उनके बीच मौजूद अंतराल का अनुमान लगाने के लिए किया जा सकता है। केवल आइडेंटिटी तथा कंजर्वेटिव रिप्लेसमेंटों को ही ध्यान में रखा जाता है।

ब्लास्ट सीक्वेंस सर्च टूल्स

तालिका 1: यहां वर्णित पांच ब्लास्ट कार्यक्रम निम्न कार्य निष्पादित करते हैंः

कार्यक्रम	क्वेरी सीक्वेंस	डेटाबेस	तुलना
ब्लास्टp	प्रोटीन	प्रोटीन	प्रोटीन
ब्लास्टn	डीएनए	डीएनए	डीएनए
ब्लास्टx	डीएनए	प्रोटीन	प्रोटीन
टीब्लास्टn	प्रोटीन	डीएनए	प्रोटीन
टीब्लास्टx	डीएनए	डीएनए	प्रोटीन

ब्लास्टचः यह विधि प्रोटीन सीक्वेंस डेटाबेस की एमिनो अम्ल क्वेरी सीक्वेंस के साथ तुलना करता है।

ब्लास्टदः यह विधि न्यूक्लियोटाइड सीक्वेंस डेटाबेस की न्यूक्लियोटाइड क्वेरी सीक्वेंस के साथ तुलना करता है।

ब्लास्टगः यह विधि प्रोटीन डेटाबेस के विरुद्ध न्यूक्लियोटाइड सीक्वेंस के छह–फ्रेम ट्रांसलेशन उत्पादों की खोज करता है।

टीब्लास्टदः यह विधि डेटाबेसों में ट्रांसलेटिड न्यूक्लियोटाइड सीक्वेंस के विरुद्ध प्रोटीन सीक्वेंस की खोज करता है।

टीब्लास्टगः यह विधि न्यूक्लियोटाइड सीक्वेंस डेटाबेस के छह–फ्रेम ट्रांसलेशनों की न्यूक्लियोटाइड क्वेरी सीक्वेंस के छह फ्रेम ट्रांसलेशनों के साथ तुलना करता है। यह कार्यक्रम ब्लास्टग तथा टीब्लास्टद कार्यक्रम के समान है।

मुख्य मानक

DATALIB: डेटाबेस या डेटाबेसों का समूह, खोज के लिए चुना गया।

MATRIX: प्रयुक्त डिस्टेंस मैट्रिक्सः ब्लोसम 62 (डिफाल्ट), पीएएम40, पीएएम120, पीएएम250 आदि।

CUTOFF: स्कोर एस. डिफाल्ट द्वारा एक्सपेक्ट से निर्धारित किया गया।

EXPECT: पाए जाने वाले अपेक्षित बेतरतीब हिटों की संख्या (डिफाल्ट = 10)

FILTER : दो में से एक 'फिल्टरों' के उपयोग का विकल्प, 'निम्न सूचना अंश' के अत्यधिक आवर्ती सीक्वेंसों या सीक्वेंसों की उपेक्षा करना।

ब्लास्ट एल्गोरिथ्म निम्न चरणों में कार्य करता हैः

क्वेरी का पूर्व संसाधन

प्रथम चरण डेटाबेस से क्वेरी सीक्वेंसों व सीक्वेंसों के बीच अंतरालहीन समानता वाले क्षेत्रों का शीघ्रता से पता लगाना है। इसी प्रकार क्वेरी के लंबाई, ू (टपल या शब्दों) की तुलना सभी डेटाबेस सीक्वेंसों से की जाती है।

ब्लास्ट में सीक्वेंस के अक्षरों से निर्मित लंबाई, ू के सभी डेटाबेस शब्दों की तुलना सृजित क्रमों के अक्षरों से की जाती है (उदाहरण के लिए एमिनो अम्ल क्रमों के साथ यदि ू=2 हो तो 20^2=400 संभावित शब्द हो सकते हैं और यदि ू=3 हो तो 20^3=8000 शब्द हो सकते हैं)। क्वेरी के प्रत्येक शब्द की तुलना इस विशाल सैट के प्रत्येक शब्द से की जाती है तथा थ्रेशहोल्ड ज का उपयोग सैट के प्रत्येक शब्द की समानता को व्यक्त करने के लिए किया जाता है। क्वेरी सीक्वेंस की प्रत्येक स्थिति का संबंध शब्दों की उस सूची से होता है जो इस स्थिति से आरंभ होने वाली क्वेरी के प्रत्येक शब्द से तब तुलनीय होता है जब उसका स्कोर ज से अधिक होता है। समान शब्द पड़ोसी कहलाते हैं।

बिटों का सृजन

मान लीजिए कि क एक डेटाबेस का सीक्वेंस है तथा फ क्वेरी सीक्वेंस है। प्रथम चरण के पश्चात् क्वरी सीक्वेंस फ को अब पड़ोसियों की सूची द्वारा अभिव्यक्त किया जाता है, क्वेरी की प्रत्येक स्थिति के लिए एक सूची। फ की क के साथ तुलना करने पर क शब्दों तथा फ की प्रत्येक स्थिति पर पड़ोसियों के बीच समानताएं दिखाई देती हैं। इसलिए फ की प्रत्येक स्थिति की तुलना क के प्रत्येक शब्द से की जाती है और यदि पड़ोसी के एक शब्द की भी स्थिति फ के समान होती है तो शब्द क जो एक हिट होता है, रिकॉर्ड किया जाता है। एक हिट समान शब्दों के एक या अनेक परवर्ती (ओवरलैपिंग) युग्मों से बनाई जाती है तथा इसका लक्षण–वर्णन दो सीक्वेंसों में प्रत्येक की स्थिति से किया जाता है। डेटाबेस से क्वेरी सीक्वेंस तथा सीक्वेंसों के बीच सभी संभावित हिटों की इस प्रकार गणना की जाती है।

बिटों का विस्तार

प्रत्येक सृजित की गई हिट का अब विस्तार किया जाता है, जिसमें कोई अंतराल नहीं रहने दिया जाता, ताकि यह पता लगाया जा सके कि यह हिट समानता के बड़े खंड का भाग हो सकता है या नहीं। इसलिए प्रत्येक हिट को दोनों दिशाओं में विस्तारित किया जाता है तथा विस्तार चरण को तेजी से सम्पन्न करने के लिए इस विस्तार को उसी समय रोक दिया जाता है जब विस्तारित हिट ग की तुलना में घटने लगता है (मान का चयन) विधि के ग मानक के लिए किया जाता है। इसके साथ ही इसकी तुलना उस सर्वश्रेष्ठ स्कोर से की जाती है जो विस्तार प्रक्रिया के दौरान प्राप्त होता है।

ब्लास्ट के प्रकार

मेगाब्लास्ट खोजः यह मौलिक ब्लास्ट खोज की तरह है लेकिन इसमें कुछ मानकों में परिवर्तन की अनुमति होती है, ताकि अधिक विशेषीकृत ब्लास्ट की खोज की जा सके।

अपनी खोज के लिए किसी जीव या वर्गीकरण विज्ञान वर्ग को विशिष्टीकृत करना

E मान सैट करना

निम्न जटिलता या ह्यूमैन रिपीट्स के लिए फिल्टर

क्वेरी आनुवंशिक कोड (ब्लास्टग तथा टीब्लास्टग केवल)

अपनी स्कोरिंग मैट्रिक्स को परिवर्तित करना

कुछ अन्य प्रगत ब्लास्ट विकल्प भी हैं।

पीएसआई–ब्लास्टः पोजीशन स्पेसिफिक इटिरेटिव–ब्लास्ट, इस युक्ति का उपयोग तब किया जा सकता है जब आपके ब्लास्ट खोज संबंधी परिणाम केवल कुछ मिलानों के साथ दिए गए हों, पीएसआई–ब्लास्ट को परिभाषित प्रोफाइल सृजित करते हुए ब्लास्ट खोजों में री–इटिरेट किया जाता है, री–इटिरेशन पर आप वे एलाइनमेंट मैच देख सकते हैं जो इस दृष्टि से उल्लेखनीय होते हैं कि उनका पता आप केवल ब्लास्ट का प्रयोग करके नहीं लगा सकते।

पीएचआई–ब्लास्टः पैटर्न हिट इनिसिएटिव ब्लास्ट, युक्ति का उपयोग आपके क्रम में विशिष्ट पैटर्न या मॉटिफ की खोज के लिए किया जा सकता है तथा इसका उपयोग जिन एमिनो अम्ल सीक्वेंस की आप खोज कर रहे हैं उनके डेटाबेसों के निर्धारित पैटर्न के लिए भी हो सकता है।

ब्लास्ट 2 सीक्वेंसः इस युक्ति से स्थानीय एलाइनमेंट के लिए ब्लास्ट इंजन का उपयोग करके किन्हीं दो क्रमों के एलाइनमेंट के लिए किया जाता है।

ब्लास्ट युक्ति में शामिल चरण

ब्लास्ट खोज किस प्रकार करें :

चरण 1: एनसीबीआई साइट पर ब्लास्ट पेज पर जाएं। ब्लास्ट से संबंधित सभी युक्तियां/कार्यक्रम एनसीबीआई से उपलब्ध हैं।

चरण 2: डेटा एंट्री फील्ड में FASTA फार्मेट सीक्वेंस पेस्ट करें।

चरण 3: सीक्वेंस एंट्री बॉक्स के नीचे अनेक बॉक्स हैं। आजमाने के उद्देश्य से मानों के डिफाल्ट के लिए सभी फील्डों को छोड़ दें।

चरण 4: महत्वपूर्ण फील्ड ड्रॉप मेन्यू है जिससे डेटाबेस के उपयोग हेतु चयन किया जा सकता है। दत (नॉन रिडन्डेंट डेटाबेस) डिफाल्ट सैटिंग है। पसंद के अनुसार सूची से किसी भी डेटाबेस को चुना जा सकता है।

चरण 5: एकमात्र विकल्प जिसे परिवर्तित किया जा सकता है, 'फिल्टरिंग' है। 'चूज फिल्टर' चैक बॉक्स को अनचैक करें। वास्तविक सकारात्मक हिटों को प्राप्त करने के लिए सीक्वेंसों को सामान्यतः फिल्टर किया जाना चाहिए। फिल्टर विकल्प यह सुनिश्चित करता है कि ऐसे छोटे सीक्वेंसों के कारण कोई मिथ्या सकारात्मक परिणाम प्राप्त नहीं होगा जो जीवविज्ञानी सीक्वेंस डेटाबेसों/स्पैक्ट्रम में बहुत सामान्य हैं।

चरण 6: अब सर्च आरंभ करने के लिए 'ब्लास्ट' बटन को क्लिक करें।

चरण 7: अब एक नया पेज दिखाई देगा जिसमें खोज की आईडी संख्या होगी तथा लगभग प्रतीक्षा समय भी प्रदर्शित होगा। 'फोर्मेट' बटन को क्लिक करें और परिणामों की प्रतीक्षा करें। जब खोज पूरी हो जाएगी तो परिणाम प्राप्त हो जाएंगे।

ब्लास्ट परिणामों की व्याख्या

1. अधिक जटिल खोज समस्याओं के लिए ब्लास्ट हिटों की लंबाई तब अधिक महत्वपूर्ण हो जाती है, जब :

   क्वेरी सीक्वेंस छोटा हो (100 न्यूक्लियोटाइडों या एमिनो अम्लों से कम लंबाई) यहां तक कि शीर्ष E- मानों का सटीक मिलान $1x\ 10^{-50}$ से अधिक हो। हमें शीर्ष हिटों की आईडेंटिटी के प्रतिशत की जांच करनी होगी, न कि केवल E–मानों की।

   निम्न E-मानों वाले वे हिट जिनमें क्वेरी सीक्वेंस के छोटे क्षेत्रों के साथ समानता मौजूद होती है, यह इंगित कर सकते हैं कि सीक्वेंस में मॉटिफ या कार्यात्मक डोमेन की समानताएं हो सकती हैं न कि वे संबंधित जीनों या प्रोटीनों का प्रतिनिधित्व करते हैं।

   उच्चतर E- मानों वाले हिट, जो $1x10^{-50}$ से $1x10^{-5}$ के बीच के होते हैं, यह इंगित करते हैं कि क्वेरी तथा हिट एक–दूसरे से संबंधित हैं। यदि हिट में क्वेरी से कम से कम 35 प्रतिशत समानता होती है तो इसकी लंबाई से भी कम से कम 80 प्रतिशत समानता होती है।

2. प्राप्त हिटों का उच्च प्रतिशत यह दर्शाता है कि मिलानों से संबंधित क्वेरी के मौजूद होने की अधिक संभावना है लेकिन यदि समानता केवल छोटे क्षेत्रों में होती है तो इससे यह संकेत मिलता है कि इसमें कार्यात्मक डोमेन मौजूद हैं, न कि संबंधित प्रोटीन/जीन उत्पाद मौजूद हैं।

शब्दावली:

पोजीटिव/नेगेटिव

यह एक विधि द्वारा उत्पन्न किया गया लेबल है, उदाहरण: एक बिंदु प्लाट में बिंदु पोजिटिव होते हैं तथा अ–बिंदु नेगेटिव होते हैं।

ट्रू/फाल्स:

ट्रू बिंदु वे हैं जो सकारात्मक (समांगी) के रूप में सही रूप से एसाइन किए जाते हैं या नकारात्मक (अ–समांगी) के रूप में सही रूप से एसाइन किए जाते हैं।

इस प्रकार चार संभावनाएं हैं।

	पोजीटिव	नेगेटिव
पोजीटिव	P^+ ट्रू पोजीटिव	N^- फाल्स पोजीटिव
नेगेटिव	P^- फाल्स नेगेटिव	N+ ट्रू नेगेटिव

सीक्वेंस एलाइनमेंट के लिए सांख्यिकी प्राचल

किसी भी एलाइनमेंट के लिए हम स्कोर की गणना कर सकते हैं जिससे एलाइनमेंट की गुणवत्ता प्रदर्शित होती है। एलाइनमेंट की सांख्यिकीय उल्लेखनीयता का पता लगाने के लिए प्रयुक्त किए जाने वाले कुछ मानक हैं :

1- P- मानः वह संभाव्यता की तुलना स्कोर पर होगी या दिए गए थ्रेशहोल्ड के ऊपर होगी। उदाहरण : सौ से अधिक स्कोर करने वाली विंडो की संभाव्यता ($PScore_{win}>=0.2*10^{-3}$)

2. E- मानः इससे इस संभावना के बारे में सूचना उपलब्ध होती है कि दिया गया क्रम एलाइनमेंट उल्लेखनीय है। किसी डेटाबेस खोज में बड़े म्.मान कटऑफ का उपयोग करने से अधिक दूर स्थित मिलानों को पाया जा सकता है लेकिन इससे गलत एलाइनमेंट भी हो सकते हैं। डेटाबेस खोज के लिए सामान्यतः 0.01 से 0.001 E मानों का उपयोग किया जाता है।

3. Z- स्कोर (मानकीकृत स्कोर, मानक सामान्य विचलन) : यह बेतरतीब एलाइनमेंट से संबंधित किसी एलाइनमेंट की उल्लेखनीयता का माप है, जो ब्लास्ट म्—मान के समान होता है।

Z= ¼Obs_score-Exp_score½@Std_deviation

नियम :

Z<-3 समांगता का कोई प्रमाण नहीं

-3<Z<6 समांगता संभव

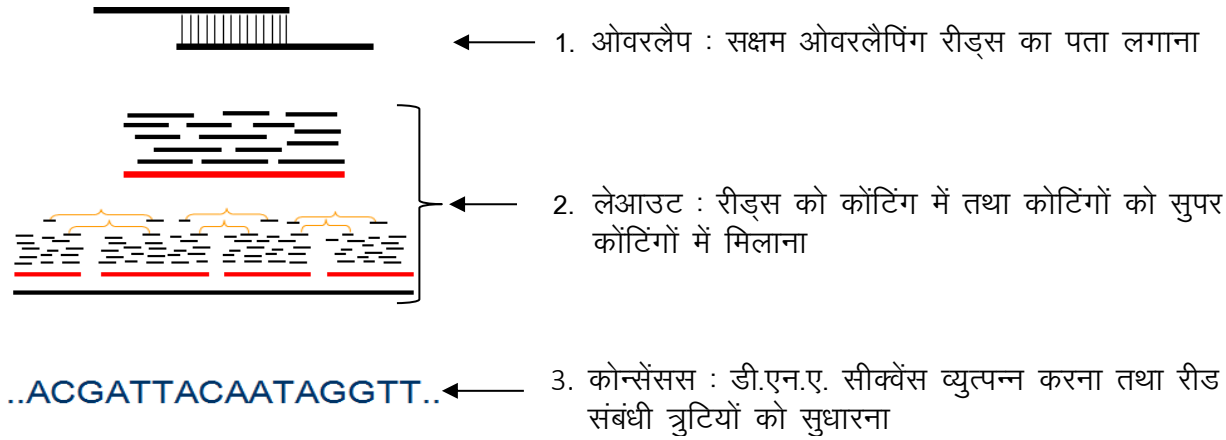
6<Z समांगता का सशक्त प्रमाण, (Z>8) बेहतर

जीनोम असेंबली

प्रस्तावना

सीक्वेंस असेम्बली का अर्थ मूल सीक्वेंस के निर्माण के लिए डी.एन.ए. सीक्वेंस के खंडों को एलाइन करना व उन्हें आपस में मिलाना है। यह अपरिहार्य है क्योंकि डी.एन.ए. सीक्वेंसिंग प्रौद्योगिकी से एक बार में सम्पूर्ण जीनोमों को नहीं पढ़ा जा सकता है बल्कि उन्हें खंडों में पढ़ा जाता है और बेतरतीब क्रम में पढ़ा जाता है जो 20 और 1000 बेसों के बीच होते हैं और यह प्रयुक्त प्रौद्योगिकी पर निर्भर करता है। सीक्वेंसिंग प्रौद्योगिकी में हुई हाल की प्रगतियों से बड़ी मात्रा में सीक्वेंस आंकड़े सृजित करना संभव हुआ है। इन उच्च – थ्रू पुट विधियों से उत्पन्न खंड, तथापि, परंपरागत सेंगर सीक्वेंसिंग विधि की तुलना में काफी छोटे होते हैं।

प्रथम सीक्वेंस एसेम्बलर 1980 के दशक के अंत में तथा 1990 के दशक के आरंभ में साधारण सीक्वेंस एलाइनमेंट कार्यक्रमों के वेरिएंट के रूप में देखे गए थे जिनमें डी.एन.ए. सीक्वेंसर कहलाने वाले स्वचालित सीक्वेंसिंग इंस्ट्रूमेंट्स द्वारा बड़ी मात्रा में खंड सृजित हुए थे। सम्पूर्ण जीनोम शॉटगन (WGS) खंड असेम्बली के लिए एल्गोरिथ्म विकसित किए गए जिनमें एटलस, एराक्ने, सेलेरा, पीसीएपी, फ्रैप (www.phrap.org) और फ्यूजन शामिल हैं। ये सभी कार्यक्रम ओवरलैप–ले आउट – कन्सेंसस एप्रोच पर आधारित होते हैं जहां सभी रीड्स की तुलना युग्मवार फैशन में एक–दूसरे से की जाती है।



चित्र 1: स्कैफोल्ड्स

परिणामस्वरूप प्राप्त (ड्राफ्ट) जीनोम सीक्वेंस क्रमबद्ध किए गए 'कोंटिग्स' की सूचना को मिलाकर किया जाता है और उसके बाद 'स्कैफोल्ड' सृजित करने के लिए संबंधित सूचना का उपयोग किया जाता है (चित्र 1)। स्कैफोल्ड 'सुनहरा पथ' सृजित करने के लिए गुणसूत्रों के भौतिक मानचित्र के साथ स्थित होते हैं।

हाल ही में एक नई अनुक्रमण विधि विकसित हुई है। वाणिज्यिक रूप से उपलब्ध प्रौद्योगिकियों में शामिल हैं : पाइरोसीक्वेंसिंग (454 सीक्वेंसिंग), संश्लेषण द्वारा सीक्वेंसिंग (इल्यूमिना) और

लाइगेशन द्वारा सीक्वेंसिंग (एसओएलआईडी)। इन अगली पीढ़ी की सीक्वेंसिंग प्रौद्योगिकियों द्वारा सृजित रीड्स परंपरागत सैंगर रीड्स की तुलना में काफी छोटे होते हैं। अपनी छोटी लंबाई के कारण इन्हें बड़ी मात्रा में उत्पन्न किया जाना चाहिए तथा पूर्व की सीक्वेंसिंग तकनीकों की तुलना में इनके द्वारा अधिक सीक्वेंसिंग डेफ्थ होनी चाहिए जबकि लंबे रीड्स से लंबे ओवरलैप उपलब्ध होते हैं जिनसे वास्तविक ओवरलैप सुस्पष्ट हो जाते हैं, रिपीट्स में छोटे रीडों के अंतर निर्धारित किए जा सकते हैं। इन मुद्दों के कारण इन अत्यंत छोटे रीड्स के लिए विशेष रूप से नई असेम्बली युक्तियां डिजाइन करने के लिए अनेक अनुसंधान दल उभर कर सामने आए हैं।

सीक्वेंसर के प्रकार और डेटा फार्मेट

इल्यूमिना	:	FASTQ
SoLID/ABI-Life	:	FASTA
Roche 454	:	SFF
Ion Torrent	:	SFF या FASTQ

असेम्बली के प्रकार

संदर्भ जीनोम की उपलब्धता के आधार पर असेम्बली के दो प्रकार हैं :

क) डी नोवो असेम्बली : रीड्स एक–दूसरे के साथ एलाइन किए जाते हैं ताकि कन्सेंसस सीक्वेंस निर्मित हो सके जो कोंटिंग कहलाते हैं।

ख) संदर्भ जीनोम एसेम्बली : यहां रीड्स एक कन्सेंसस सीक्वेंस का निर्माण करने के लिए उपलब्ध संदर्भ जीनोम से एलाइन किए जाते हैं।

जीनोम असेम्बली की तकनीकें

लगभग सभी बड़े पैमाने की सीक्वेंसिंग परियोजनाओं में ऐसी शॉटगन कार्यनीति का उपयोग होता है जिसमें एसेम्बलर (डेड्यूस) लक्षित क्रम से छोटे डी.एन.ए. खंडों के सैट से डी.एन.ए. सीक्वेंस को लक्षित करता है। छोटे डी.एन.ए. खंडों का सैट जो शॉटगन रीड्स कहलाता है, कोंटिंग या एलाइंड खण्डों के सैट के रूप में असेम्बल किया जाता है जिसके लिए फ्रेग्मेंट असेम्बलर नामक कलन विधि का उपयोग होता है। फ्रेग्मेंट असेम्बली एक संकल्पनात्मक सरल प्रक्रिया है जिसमें ओवरलैपिंग खण्डों की पहचान करके अपेक्षाकृत लंबे सीक्वेंस सृजित किए जाते हैं। यदि खण्ड एसेम्बली को सटीकता से किया जाए तो जीनोम सीक्वेंसिंग की समस्या सरल हो जाती है। तथापि, पुनरावृत्ति सीक्वेंस होते हैं जो अल्पकाल में पुनरावृत्ति होते हैं तथा जीनोमी क्रम में रहते हैं जिनसे खण्ड असेम्बली की प्रक्रिया में बहुत आसानी से गलती हो सकती है। रिपीट्स से उत्पन्न होने वाली कठिनाई को दूर करने के लिए उपयोगी तकनीक यह है कि क्लोन के दोनों छोरों को सीक्वेंस किया जाए जिससे प्रतिक्लोन दो खण्ड रीड सृजित हों। चूंकि क्लोन का इन्सर्ट आकार ज्ञात होता है अतः हम दो खंडों के बीच की लगभग दूरी जानते हैं। खण्ड मिलान संबंधी सूचना भी अक्सर मेट–पेयर सूचना कही जाती है जो बड़े पैमाने पर शॉटगन सीक्वेंसिंग के लिए अनिवार्य हो जाती है। असेम्बली प्रक्रिया के

दौरान इस सूचना के उपयोग में मुख्य मुद्दा यह है कि हम दो रीड्स के बीच के क्रम को नहीं जानते हैं और इसे केवल एकल कोंटिंग में अन्य खंडों की असेम्बली द्वारा ही ज्ञात किया जा सकता है। इसलिए हम क्लोन – लंबाई संबंधी सूचना का उपयोग केवल असेम्बली के पश्चात कर सकते हैं जिससे क्लोन– लंबाई की सूचना के आधार पर सही या गलत, दोनों प्रकार की असेम्बली हो सकती है। मेट–पेयर सूचना के प्रभावी उपयोग की एक कार्यनीति सक्षम मिसअसेम्बल कोंटिंग्स का पता लगाकर यथासंभव सटीक कोंटिंग्स को असेम्बल करना तथा इसके बाद सही रूप से असेम्बल किए गए कोंटिंग्स का ही उपयोग करके मेट–पेयर सूचना का इस्तेमाल करना है। जीनोम–सीक्वेंसिंग केन्द्रों में जीनोम सीक्वेंसिंग तथा असेम्बली पर बल देने के लिए इस्तेमाल होने वाली सामान्य प्रक्रिया है :

1. खण्ड रीडआउटः प्रत्येक खण्ड का सीक्वेंस स्वचालित बेस–कालिंग सॉफ्टवेयर का उपयोग करके पता लगाया जाता है। फ्रैड सर्वाधिक व्यापक रूप से प्रयुक्त होने वाला कलन विधि है।
2. वाहक सीक्वेंस को कतरनाः शॉटगन रीड्स में वाहक क्रमों का वह भाग होता है जिसे सीक्वेंस असेम्बली के पूर्व हटाना होता है।
3. निम्न गुणवत्ता वाले क्रमों को कतरनाः शॉटगन रीड्स में घटिया गुणवत्ता वाले बेस काल होते हैं और इन निम्न गुणवत्ता वाले बेस काल को हटाने या उन्हें ढक देने से अक्सर अधिक सटीक सीक्वेंस असेम्बली होती है। तथापि यह चरण वैकल्पिक है तथा सीक्वेंसिंग करने वाले कुछ केन्द्र निम्न गुणवत्ता वाले बेस कालों को ढकते नहीं हैं तथा सच्चे खण्ड ओवरलैपों के बारे में निर्णय लेने के लिए गुणवत्तापूर्ण मानों के उपयोग की दृष्टि से फ्रेगमेंट असेम्बलर पर निर्भर रहते हैं।
4. खण्ड असेम्बलीः शॉटगन डेटा उस खण्ड असेम्बलर के लिए इनपुट है जो कोंटिंग्स कहलाने वाले एलाइन किए गए खण्ड के सैट को स्वतः ही सृजित करता है।
5. असेम्बली सत्यापनः पिछले चरणों में असेम्बल किए गए कुछ कोंटिंग रिपीट के कारण मिसअसेम्बल हो जाते हैं। चूंकि हमें लक्ष्य डी.एन.ए. में रिपीटों की पूर्व जानकारी नहीं होती है अतः प्रत्येक कोंटिंग में असेम्बली के सही होने को सत्यापित करना बहुत कठिन है और यह चरण अधिकांशतः मानवीय विधि से सम्पन्न किया जाता है। कोंटिंग असेम्बलियों के स्वचालित सत्यापन से संबंधित हाल ही में कुछ एल्गोरिद्मिक विकास हुए हैं।
6. स्कैफोल्डिंग कोंटिगः कोंटिंग अभिमुख तथा क्रमबद्ध होने चाहिए। मेट–पेयर सूचना इस चरण के लिए प्राथमिक सूचना है, अतः यदि इनपुट शॉटगन को क्लोनों के दोनों छोरों की रीडिंग द्वारा तैयार नहीं किया जाता है तो यह चरण पूरा नहीं हो सकता है।
7. समाप्तिः यह ज्ञात करने के लिए कि सभी कोंटिंग ठीक से असेम्बल हुए हैं और कोंटिंग अभिमुखित हैं व सही क्रम में हैं, हम अंतरालों की स्थिति से सम्बद्ध सीक्वेंसिंग विशिष्ट क्षेत्रों द्वारा दो कोंटिंग के बीच के अंतराल को भर सकते हैं।

अगली पीढ़ी के सीक्वेंसिंग (एनजीएस) रीड्स की नवीन असेम्बली

एनजीएस रीड सृजित होने के पश्चात् उन्हें ज्ञात संदर्भ सीक्वेंस के रूप में एलाइन किया जाता है या नवीन रूप से असेम्बल किया जाता है। नवीन असेम्बली जीवों के जीनोम की

पुनर्संरचना की प्रक्रिया है। ये जीव इससे पहले सीक्वेंस नहीं किए गए होते हैं या इनके संदर्भ तुलनात्मक जीनोम उपलब्ध नहीं होते हैं। इसे शॉटगन प्रक्रिया से सम्पन्न किया जाता है जहां जीव के जीनोम को छोटे खण्डों में विभाजित किया जाता है और प्रत्येक को अलग–अलग सीक्वेंस किया जाता है या कम्प्यूटेशनल युक्तियों का उपयोग करके उन्हें पुनः निर्मित किया जाता है। यह प्रक्रिया जटिल है क्योंकि जीनोम में समरूप सीक्वेंस के खण्ड होते हैं जो रिपीट कहलाते हैं। रिपीटों की लंबाई अत्यधिक भिन्न होती है जिससे सम्पूर्ण जीनोम को प्राप्त करना असंभव हो जाता है। इसलिए, लगभग सभी नई युक्तियों से पूर्ण जीनोम को प्राप्त नहीं किया जा सकता है। तथापि, इनमें कोंटिंग के नाम से ज्ञात जीनोम के लंबे खण्ड होते हैं। इसके अतिरिक्त जीनोम के आकार के बढ़ने के साथ जटिलता भी बढ़ती जाती है। नवीन जीनोम असेम्बली की प्रक्रिया में प्राथमिकतः दो श्रेणियां होती हैं, नामतः ओवरलैप लेआउट एंड कोंसेंसस (ओएलसी) तथा डी ब्रूजिन ग्राफ आधारित विधि। इनमें से पहली विधि मैमोरी गहन है। डी ब्रूजिन ग्राफ पर आधारित अनेक युक्तियां उपलब्ध हैं।

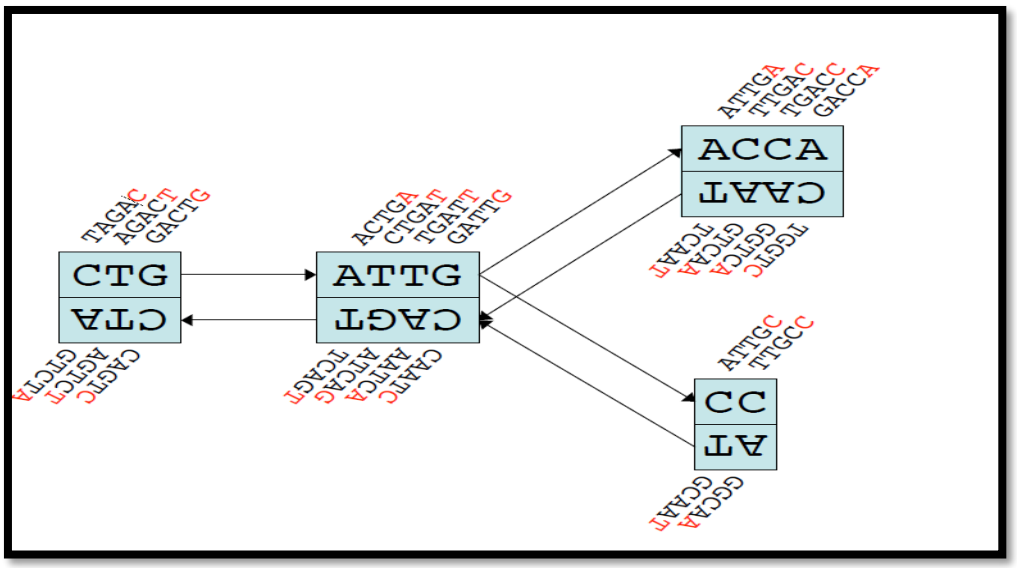
दोहरे छोर वाले शॉर्ट–रीड सीक्वेंसिंग प्रौद्योगिकियों के लिए असेम्बली

हाल ही में विकसित पायरोसीक्वेंसिंग – जैसी तकनीकें अत्यधिक आशाजनक हैं। तथापि, इनमें परिणामस्वरूप प्राप्त रीड्स की लंबाई वर्तमान सीक्वेंसिंग मशीनों द्वारा उत्पन्न रीड्स की तुलना में अत्यधिक कम होती है। सीक्वेंस रिपीट की लंबाई रीड्स की उपयोगिता को सीमित कर देती है। क्योंकि जब किसी भी सीक्वेंस रिपीट की लंबाई बढ़ जाती है तो इसे गैर समाधान के रूप में परिभाषित किया जाता है। विशेष रूप से छोटे रीड्स के संकलन में सबसे छोटा सामान्य सुपरस्ट्रिंग लक्ष्य के अत्यधिक सम्पीड़ित आकार का प्रतिनिधित्व करता है। रिपीटों की इस समस्या अर्थात् भिन्नतापूर्ण इंसर्ट लंबाई, को हल करने के लिए दोहरे छोर रीड वाले प्रोटोकाल प्रस्तावित किए गए। खण्ड–बहु लक्षित क्लोन तथा लंबाई a±b:, या (समतुल्य बताए गए) d से d+w लंबाई के सभी खण्डों को अलग करने के लिए जैल इलेक्ट्रोफोरेसिस का उपयोग होता है। उपरोक्त समीकरण में इंटीजर d और w के मान निर्धारित होते हैं।

डी ब्रूजिन ग्राफ

वर्ष 1995 में, इड्यूरी और वाटरमैन ने असेम्बली को दर्शाने के लिए ग्राफ का उपयोग करना आरंभ किया। उन्होंने एक वैकल्पिक सीक्वेंसिंग तकनीक के लिए असेम्बली एल्गोरिथ्म प्रस्तुत किया जो संकरीकरण द्वारा सीक्वेंस को दर्शाता था जहां ओलिगोऐरे को k न्यूक्लियोटाइड संबंधी सभी शब्दों का पता लगाने के लिए इस्तेमाल किया जाता था और इन्हें ।. मर्स भी कहा जाता था जो किसी दिए गए जीनोम में उपस्थित थे। उनकी इस विधि में प्रत्येक पहचाने गए शब्द के लिए एक नोड सृजित करना और उसके बाद संबंधित नोडों को ओवरलैपिंग सम्बद्ध k मर्स से जोड़ना था। इसके बाद वे ओवरलैपिंग k- मर्स की श्रृंखला को रिपोर्ट कर सके जिससे सुस्पष्ट कोंटिंग उत्पन्न हुए क्योंकि इसमें शाखित कनेक्शन नहीं थे। यह सीक्वेंस ग्राफ डी ब्रूजिन ग्राफ कहलाता है जिसके द्वारा k–मर्स को कगारों के रूप में दर्शाया जाता है तथा ओवरलैपिंग ।.मर्स अपने छोरों से जुड़े हुए होते हैं। इसमें ग्राफ निर्माण, त्रुटि को दूर करने, मिश्रित लंबाई की असेम्बली व युग्मित–छोर की असेम्बली के लिए नए एल्गोरिथ्म होते हैं। तथापि, यह कार्यक्रम पुष्टता तथा आसानी से चलाने के लिए डिजाइन

किया गया था। इसमें कुछ विशेष पहलू हैं : पहला, यह k- मर्स को कगारो की बजाय नोड्स पर मानचित्रित करता है। दूसरा, यह विलोम सम्पूरक सीक्वेंस को बाइ–ग्राफ (या–द्विदिशा वाले ग्राफ) को प्राप्त करने के लिए क्रमबद्ध करता है अर्थात् दूसरे शब्दों में ऐसा ग्राफ है जहां एक कोर इसके किसी भी छोर पर नोड में स्वतंत्र रूप से प्रविष्टि करती है या बाहर निकलती है। प्रत्येक नोड छ, ओवरलैपिंग k-मर्स की श्रृंखला को दर्शाती है। पास के k-मर्स, k-1 न्यूक्लियोटाइडों द्वारा ओवरलैप होते हैं। k-मर्स द्वारा उपलब्ध कराई गई सीमांत सूचना इसके अंतिम न्यूक्लियोटाइड में होती है। इन अंतिम न्यूक्लियोटाइड का सीक्वेंस नोड या `(N) का सीक्वेंस कहलाता है। इसलिए नोड का सीक्वेंस सम्बद्ध k-मर्स को अपूर्ण रूप से दर्शाता है। दूसरे शब्दों में k-मर्स के दो अलग–अलग सैट समान सीक्वेंस से युक्त दो अलग–अलग नोडों द्वारा दर्शाए जा सकते हैं। समान सीक्वेंस होने के बावजूद ये दोनों नोड अलग रखे जाते हैं तथा रीड्स का मानचित्रण उनमें निहित ।.मर्स के अनुसार किया जाता है। प्रत्येक नोड छ जुड़वां नोड ~N से जुड़ा होता है जो विलोम पूरक k-मर्स की विलोम श्रृंखला को दर्शाता है। इससे यह सुनिश्चित होता है कि विपरीत लड़ियों से लिए गए रीड्स जो एक–दूसरे को ओवरलैप करते हैं, उन्हें प्रयोग में शामिल कर लिया गया है। यह ध्यान देना महत्वपूर्ण है कि नोड या इसके जुड़वां के साथ सम्बद्ध सीक्वेंस एक–दूसरे के विलोम रूप से पूरक हों, यह आवश्यक नहीं है। नोड N तथा इसके जोड़े का मेल ब्लॉक कहलाता है। इसके पश्चात नोड में होने वाला कोई भी परिवर्तन इसके जोड़े के लिए भी समान रूप से लागू होता है। ब्लॉकों को इम्पलीसिट बाई–ग्राफ के नोड के रूप में माना जा सकता है। नोडों को एक निर्देशित कोर या चाप द्वारा जोड़ा जा सकता है। ऐसे मामले में किसी चाप के मूल नोड का अंतिम k-मर प्रथम डेस्टिनेशन नोड को ओवरलैप करता है। ब्लॉकों में सममितीय होने के कारण यदि चाप A से B नोड तक जाता है तो सममितीय ~B से ~A तक जाती है। किसी एक चाप में सुधार का अर्थ है कि इसके युग्म चाप में भी सममिति उत्पन्न होगी।



चित्र 2: डी ब्रूजिन ग्राफ के कार्यान्वयन का रेखाचित्र

एकल चतुर्भुज द्वारा दर्शाया गया प्रत्येक नोड सीधे ऊपर या नीचे सूचीबद्ध किए गए ओवरलैपिंग k-मर्स का प्रतिनिधित्व करता है (इस मामले में, k=5) प्रत्येक k-.मर का अंतिम न्यूक्लियोटाइड लाल रंग से दर्शाया गया है। चतुर्भुजों में बड़े अक्षरों में कॉपी किया गया अंतिम न्यूक्लियोटाइडों का सीक्वेंस नोड का सीक्वेंस है। नोड के नीचे या ऊपर सीधे जुड़े हुए युग्म नोड विलोम प्रतिपूरक ।.मर्स के विलोम श्रृंखला का प्रतिनिधित्व करते हैं। चापों को नोडों के बीच तीरों द्वारा दर्शाया गया है। चाप मूल का अंतिम ।.मर अपने गंतव्य के प्रथम स्थान पर ओवरलैप करता है। प्रत्येक में सममितीय चाप होता है। बाएं ओर के दो नोडों को सूचना में बिना किसी क्षति के एक साथ मिलाया जा सकता है क्योंकि ये एक श्रृंखला का निर्माण करते हैं।

- डी ब्रुजिन ग्राफ में ग्राफ के आर–पार पथों पर एक के साथ एक क्रमों के मानचित्र होते हैं। पथ से न्यूक्लियोटाइड सीक्वेंस का निष्कर्षण बिल्कुल सीधा होता है जो प्रथम नोड के आरंभिक ।.मर के रूप में दिया जाता है तथा इसे पथ के सभी नोडों में क्रमों के रूप में व्यक्त किया जाता है। प्रत्येक रीड के लिए ठीक एक पथ विद्यमान होता है जो सीक्वेंस के ।.मर्स से सम्बद्ध नोडों के माध्यम से क्रमबद्ध ढंग से आगे जाता है।
- दो ओवरलैपिंग सीक्वेंस दो पथों द्वारा अभिव्यक्त होते हैं जो एक–दूसरे को ओवरलैप करते हैं। पथों का परस्पर काट सीक्वेंसों के बीच के ओवरलैप से सम्बद्ध होता है। दो पथ टोपोलॉजी का उपग्राफ बनाते हैं जो सीक्वेंसों के बीच एलाइनमेंट के प्रकार से सीधे–सीधे जुड़ा हुआ होता है। यदि एक सीक्वेंस दूसरे को काट रहा होता है तो पथ भी अन्य पथ का उप–पथ हो जाता है। जब और क्रमों को जोड़ा जाता है तो उपरोक्त गुण प्रमाणित रहते हैं। इसका अर्थ है कि वे सभी क्रम जो समान सब्स्ट्रिंक की भागीदारी करते हैं वे रिपीटों के माध्यम से ओवरलैपिंग रीडों के सैटों की खोज करने में उपयोगी सिद्ध होते हैं क्योंकि ये एक ही पथ अपनाते हैं।

डी ब्रुजिन ग्राफ का प्रथम परिणाम यह है कि इसमें अत्यधिक विभिन्न लंबाइयों वाले सीक्वेंस को भी समायोजित किया जा सकता है। यह विशेष रूप से तब उपयोगी है जब मिश्रित लंबाई की सीक्वेंसिंग की जाती है या तुलनात्मक जीनोमिक्स में भी यह विशेष रूप से उपयोगी है। छोटे रीडों, लंबे रीडों, पूर्व एसेम्बल किए गए कोंटिग्स या अंतिम जीनोमों के लिए कोई भी तदर्थ अनुमान नहीं लगाना होता है। इसके अतिरिक्त पथ और सीक्वेंसों के बीच एक संबंध होने के कारण ओवरलैपिंग सीक्वेंस अनिवार्य रूप से समान पथ का अनुसरण करते हैं। इससे रीड्स के ओवरलैपिंग सैटों की निरंतरता के लिए खोज करना आसान हो जाता है।

जटिल जीनोमों को असेम्बल करने से संबंधित मुद्दे और इससे जुड़ी समस्याएं

जीनोम असेम्बली एक अत्यंत कठिन कम्प्यूटेशनल समस्या है जो रिड्स की अधिक संख्या तथा समान सीक्वेंसों के कारण जो रिपीट कहलाते हैं, और भी जटिल हो जाती है। यह रिपीट हजारों न्यूक्लियोटाइड लंबे हो सकते हैं और कुछ विभिन्न स्थानों पर हजारों में होते हैं, विशेष रूप से पौधों और पशुओं के बड़े जीनोमों में तो ऐसा होता ही है।

फसल जीनोमों के अनुक्रमण के मामले में एक चुनौती जीनोमों के आकार तथा विभिन्न सीक्वेंसिंग विधियों द्वारा उत्पन्न किए गए रीड्स की लंबाई में अत्यधिक अंतर है। द्वितीय पीढ़ी की सीक्वेंसिंग और आधुनिक सैंगर सीक्वेंसिंग विधि से उत्पन्न किए गए छोटे रीडों के

बीच पैमाने में 10–500 x का अंतर होता है। अनुक्रमित जीव जैसे–जैसे आकार में बढ़ते हैं, वैसे–वैसे असेम्बली कार्यक्रमों की जटिलता बढ़ती जाती है तथा जीनोम परियोजनाओं में इन समस्याओं को हल करने के लिए अत्याधुनिक कार्यनीतियां अपनाने की आवश्यकता होती है:

- टैराबाइट सीक्वेंसिंग डाटा जिन्हें कम्प्यूटिंग क्लस्टरों पर ही संसाधित कर सकते हैं;
- समरूप और लगभग समरूप क्रम (जो रिपीट्स कहलाते हैं) जो सबसे खराब अवस्था हो सकती है। एल्गोरिथ्म की समय व अंतराल की जटिलता बढ़ने के साथ–साथ चरघातांकी रूप से बढ़ते जाते हैं; और
- सीक्वेंसिंग उपकरणों से फ्रेगमेंट में होने वाली त्रुटि जो असेम्बली को परिबद्ध कर सकती है।

सारणी: विद्यमान नवीन एसेम्बलर्स की सूची

नाम	प्रकार	प्रौद्योगिकियां	लेखक	कब अद्यतन हुआ
BySS	(बड़ा) जीनोम	सोलेक्सा, सोलिड	सिम्प्सन, जे और साथी	2008 / 2011
ALLPATHS-LG	(बड़ा) जीनोम	सोलेक्सा, सोलिड	ग्नेरे, एस और साथी	2011
AMOS	जीनोम	सेंगर, 454	साल्जबर्ग, ए. और साथी	2002 / 2008
एरापन–एम	मध्यम जीनोम (जैसे ई. कोलाई)	सभी	साहली, एम. और शिबुया , टी.	2011–2012
एरापन–एस	छोटे जीनोम (विषाणु और जीवाणु)	सभी	साहली, एम. और शिबुया, टी.	2011–2012
सेलेरा डब्ल्यूजीए असेम्बलर / सीएबीओजी	(बड़ा) जीनोम	सैंगर, 454, सोलेक्सा	मायर्स, जी. और साथी; मिलर जी. और साथी	2004 / 2010
सीएलसी / जीनोमिक्स वर्कबैंच और सीएलसी असेम्बली सैल	जीनोम	सैंगर, 454, सोलेक्सा, सोलिड	सीएलसी बायो	2008 / 2010 / 2011
कॉर्टेक्स	जीनोम	सोलेक्सा, सोलिड	इकबाल, जैड और साथी	2011

डी.एन.ए. बेसर	जीनोम	सैंगर, 454	हैराकल, बायोसॉफ़्ट एसआरएल	2013
डी.एन.ए. ड्रैगन	जीनोम	इल्युमिना, सोलिड, पूर्ण जीनोमिक्स, 454, सैंगर	सीक्वेंटी एक्स	2011
डीएनएनैक्सस	जीनोम	इल्युमिना, सोलिड, पूर्ण जीनोमिक्स	डीएनएनैक्सस	2011
एडेना	जीनोम	इल्युमिना	डी. हर्नांडेज, पी. फ्रांकोइस, एल. फैरिनेली, एम. ओस्टेरॉस और जे. स्क्रैजेंल	2008 / 2013
इयूलर	जीनोम	सैंगर, 454 (सोलेक्सा)	पैवजेनेर, पी. और साथी	2001 / 2006
इयूलर–एस.आर.	जीनोम	454, सोलेक्सा	चेइसन, एमजे और साथी	2008
फोर्ज	(बड़ा) जीनोम, ईएसटी, मैटाजीनोमस	454, सोलेक्सा, सोलिड, सैंगर	प्लाट, डीएम, एवर्स, डी.	2010

संदर्भः

1. बैट्जोगलोउ, एस., जैफे, डी.बी., स्टेंले, के, बटलर, जे, ग्नेरे, एस, माउसेली, ई. बर्जर, बी, मेसिरोव, जे.पी. और साथी (जनवरी 2002). 'एराक्ने' : एक होल जीनोम शॉटगन असेम्बलर' जीनोम रिसर्च 12(1):177–89. डीओआईः10.11.1 / जीआर 208902.पीएमसी 15525. पीएमआईडी 11779843.

2. बोइसवर्ट, सेबास्टियन, लेवियोलेटे, फ्रांकोइस, कोरबेइल, जैक्स (2010). 'रेः साइमलटेनियस एसेम्बली ऑफ रीड्स फ्राम , मिक्स ऑफ हाइ थ्रू पुट सीक्वेंसिंग टैक्नोलॉजीस'. जर्नल ऑफ कम्प्यूटेशनल बायोलॉजी. 17(11) : 1519–3. डीओआईरु10. 1089 / सीएमबी.2009. 0238.पीएमसी 3119603.पीएमआईडी 20958248.

3. दोह्म, जे.सी., लोटाज. सी.; बोरोडीना, टीत्र हिमेलबाउएर, एच. (नवम्बर 2007). 'सीएचएआरसीजीएस, ए फास्ट एंड हाइली एक्यूरेट शॉर्ट रीड असेम्बली एल्गोरिथ्म फॉर

डीनोवो जीनोमिक सीक्वेंसिंग'. जीनोम रिसर्च 17(11): 1697–706. डीओआई: 10. 1101 / जीआर.6435207. पीएमसी 2045152. पीएमआईडी 17908823.

4. ह्यूस. एस.एम., ह्यूबर, जे.ए., मॉरीसन, एच.जी.,सोगिन, एम.एल. और वैल्च (डीएम) (2007). एक्यूरेसी एंड क्वालिटी ऑफ मैसिवली पैरलल डी.एन.ए. पाइरोसीक्वेंसिंग, जीनोम बायोल 8, आर 143.

5. मार्डिस, ई.आर. (2008). द इम्पेक्ट ऑफ नेक्स्ट जेनरेशन सीक्वेंसिंग टैक्नोलॉजी ऑन जेनेटिक्स, ट्रैंड्स जेनेट 24, 133–141.

6. माइकल सी. स्कार्टज, जान विटकोवस्की और डब्ल्यू रिचर्ड मैक कौम्बी (2012). करेंट चैलेंजिस इन डी नोवो प्लांट जीनोम सीक्वेंसिंग एंड असेम्बली. जीनोम बायोलॉजी, 13: 243.

7. मायर्स, ई. डब्ल्यू., सुटन, जीजी, डैल्चर, एएल, ड्यू, आई.एम., फासुले, डीपी, फलैनिगन, एमजे, क्राविड्स, एस.ए., मोवेरी, सीएम और साथी (मार्च 2000). 'ए होल जीनोम असेम्बली ऑफ ड्रोसोफिला' साइंस 287 (5461) : 2196–204. डीओआई: 10.01126 / साइंस. 287. 5461.2196. पीएमआईडी 1073113.

8. पॉप, एम. (2004) शॉटगन सीक्वेंस असेम्बली, एडवांस कम्प्यूटेशन 60, 193–248.7.

9. पॉप, एम. और स्लाजबर्ग, एस.एल. (2008). बायोइन्फोर्मेटिक्स चैलेंजिस ऑफ न्यू सीक्वेंसिंग टैक्नोलॉजी, ट्रैंड्स जेनेट 24, 142–149.

10. रोनागी, एम. उहलेन, एम और नाइरेन, पी. (1998). ए सीक्वेंसिंग मैथड बेस्ड ऑन रियल टाइम पाइरोफास्फेट, साइंस 281, 363–365.

11. झांग, डब्ल्यू, चैन जे., वांग वाई, तांग वाई, शांग जे, और साथी (2011). प्रैक्टिकल कैम्पेरीजन ऑफ डीनोवो जीनोम असेम्बली सॉफ्टवेयर टूल्स फॉर नैक्स्ट जेनेरेशन सीक्वेंसिंग टैक्नोलॉजिस. PLoS ONE 6(3):el7915.doi:10.1371/journal.pone.0017915.

जीनोम एनोटेशन

प्रस्तावना

जीनोम क्रांति होने तक अनुसंधानकर्ताओं द्वारा किसी विशेष प्रोटीन या कोशिकीय प्रक्रिया में विशिष्ट रूचि होने के कारण जीनों की पहचान की जाती थी। एक बार पहचाने जाने के बाद इन्हें आवष्यकता कैटलॉग सीक्वेंस किया गया, ऐसा विशेष रूप से क्लोनीकरण तथा cDNAs की सीक्वेंसिंग द्वारा किया गया जिसके बाद उन लंबे जीनोमिक्स खण्डों के लक्षित सीक्वेंसिंग को अपनाया गया जो cDNAs के लिए कोड किए जाते हैं। एक बार जब किसी जीव का सम्पूर्ण जीनोम सीक्वेंस उपलब्ध हो जाता है तो किसी जीनोम द्वारा कोडित सभी जीनों का पता लगाने की प्रबल होती है। इस प्रकार का केटलॉग अनुसंधानकर्ताओं के लिए अत्यंत मूल्यवान है क्योंकि जीनों की सीमित सैट की तुलना में सम्पूर्ण जानकारी से ज्यादा सीखा जा सकता है। उदाहरण के लिए समान जीनों को पहचाना जा सकता है, उनके बीच विकासात्मक तथा कार्यात्मक संबंध ज्ञात किया जा सकता है और इस आशय का एक वैश्विक चित्र प्राप्त होता है कि किसी जीनोम में कितने और किन–किन प्रकार के जीन हो सकते हैं। जीनोम सीक्वेंसिंग में प्रयासों का एक उल्लेखनीय अंग *एनोटेशन* प्रक्रिया को समर्पित है जिसमें जीन नियामक तत्व तथा सीक्वेंस के अन्य गुण, यथासंभव व्यापक रूप से पहचाने जाते हैं तथा सार्वजनिक डेटाबेसों में मानक फार्मेट में कैटलॉग किए जाते हैं ताकि अनुसंधानकर्ता सूचना का आसानी से उपयोग कर सकें। पिछले एक दशक के दौरान कार्यात्मक जीनोमिक्स अनुसंधान में अत्यधिक विस्तार हुआ है और विशेष रूप से पादप जीवविज्ञान अनुसंधान समुदाय ने इस दिशा में बहुत कार्य किया है। नवीन डी.एन.ए. सीक्वेंसों का कार्यात्मक एनोटेशन कार्यात्मक जीनोमिक्स में शीर्ष आवश्यकता वाला है और काफी हद तक प्रायोगिक परिणामों की जीवविज्ञानी व्याख्या में मुख्य भूमिका निभाता है।

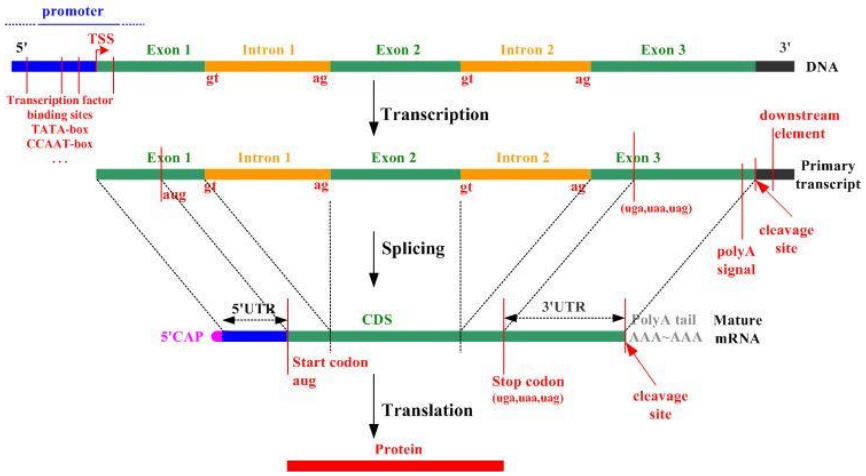
कम्प्यूटेशनल जीन पूर्वानुमान

कम्प्यूटेशनल जीन पूर्वानुमान स्वचालित विश्लेषण में अधिक से अधिक अनिवार्य होता जा रहा है और बड़े सीक्वेंसों के एनोटेशन में भी प्रतिदिन महत्वपूर्ण हो रहा है। पिछले दो दशकों में डी.एन.ए. सीक्वेंसों के प्रोटीन कोडीकरण क्षेत्रों के पूर्वानुमान के लिए अनेक एल्गोरिथ्म विकसित किए गए हैं। ये काफी हद तक एक–दूसरे के समान हैं जिनमें एक्सॉन, इंट्रॉन, स्प्लाइसिंग स्थलों, रेगुलेटरी स्थलों आदि जैसे जीन संबंधी गुणों में भेद करने की क्षमता है। जीन पूर्वानुमान विधियों से क्वेरी सीक्वेंसों में पहले कोडिंग क्षेत्र का पूर्वानुमान लगाया जाता है और उसके बाद सीक्वेंस डेटाबेसों को एनोटेट किया जाता है।

जीन संरचना और अभिव्यक्ति

यूकैरियाटों में जीन संरचना और जीन अभिव्यक्ति प्रोकैरियाटों की तुलना में अधिक जटिल है। विशिष्ट यूकैरियाटों में किसी प्रोटीन के लिए डी.एन.ए. कोडिंग का क्षेत्र सामान्यतः निरंतर नहीं होता है। यह क्षेत्र *एक्सॉन* और *इंट्रॉन* के एकांतरिक स्ट्रैच से बना होता है। ट्रांसक्रिप्सन के दौरान एक्सॉन और इंट्रॉन दोनों आरएनए पर ट्रांसक्राइब होते हैं और उनका रैखिक क्रम होता है। इसके पश्चात् *स्पलाइसिंग* कहलाने वाली प्रक्रिया होती है जिसमें इंट्रॉन सीक्वेंस खंडित हो जाता है तथा आरएनए क्रम से बहिष्कृत हो जाता है। शेष आरएनए खण्ड जब एक बार संबंधित एक्सॉनों के साथ मिलकर परिपक्व आरएनए का निर्माण करते हैं तो

आरएनए लड़ी तैयार होती है। किसी विशिष्ट बहुएक्सॉन वाले जीन की संरचना उपरोक्त चित्र 1 में दर्शायी गई है। यह प्रमोटर क्षेत्र से आरंभ होती है जो बाद में ट्रांसक्राइब लेकिन अ–कोडीकरण क्षेत्र के पूर्व आती है। इस क्षेत्र को *5´ अनट्रांसलेटिड रीजन (5´ यूटीआर)* कहते हैं। इसके पश्चात् आरंभिक एक्सॉन आता है जिसमें स्टार्ट कोडोन होता है। आरंभिक एक्सॉन के पश्चात् इंट्रॉन और आंतरिक एक्सॉनों की एकांतरिक श्रृंखला होती है जिसके पश्चात् टर्मिनेटिंग एक्सॉन आता है जिसमें स्टॉप कोडोन होता है। इसके बाद एक अन्य अ–कोडीकरण क्षेत्र आता है जिसे *3´ यूटीआर* कहते हैं। यूकैरियोटी जीन के अंत में एक पॉलीएडेनाइलेशन (पॉली ए) संकेत होता है : न्यूक्लियोटाइड एडेनीन जो कई बार रिपीट होता है, एक्सॉन–इंट्रॉन सीमाओं (जैसे स्प्लाइस स्थलों) को विशिष्ट छोटे (2इच लंबे) सीक्वेंसों द्वारा संकेत किया जाता है। किसी इंट्रॉन (एक्सॉन) का 5′(3′) छोर *डोनर* स्थल कहलाता है तथा इंट्रॉन (एक्सॉन) 3′(5′) छोर *एक्सेप्टर* स्थल कहलाता है। यूकैरियाटों के मामले में जीन पहचान की समस्या जटिल हो जाती है क्योंकि इनकी जीन संरचना में अत्यंत विविधता पाई जाती है।



चित्र 1. प्रोटीन कोडिंग यूकैरियाटीक जीन का रेखाचित्र

जीन पूर्वानुमान की विधियां

कम्प्यूटेशनल जीन पूर्वानुमान की विधियों के दो मुख्य वर्ग हैं (चित्र 2)। एक सीक्वेंस की समानता पर आधारित है, जबकि दूसरा जीन संरचना तथा संकेत पर आधारित विधियां हैं जिन्हें ।इ इनिशियो जीन फाइंडिंग कहा जाता है।

सीक्वेंस समानता संबंधी खोजें

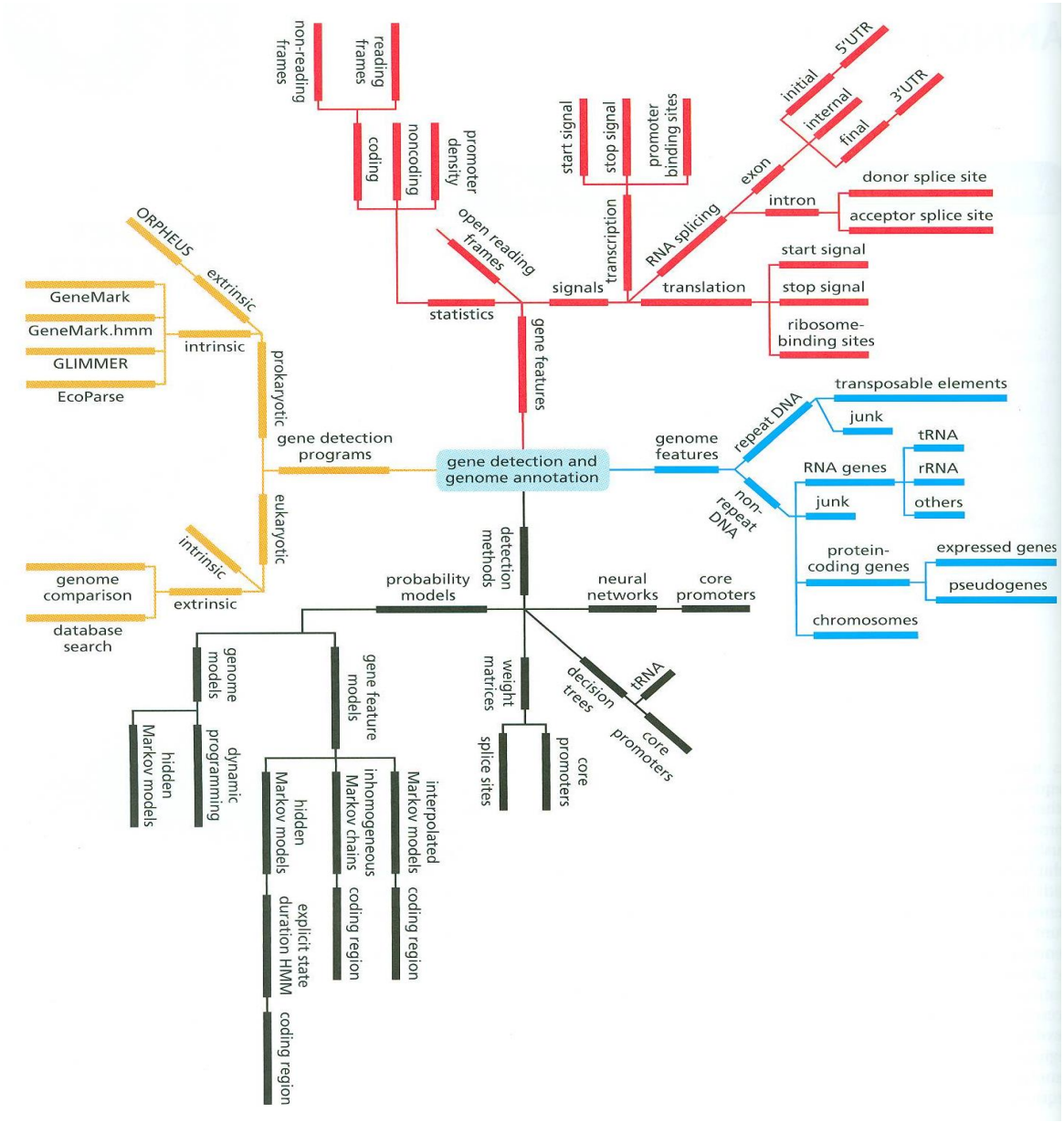
सीक्वेंस समानता संबंधी खोज संकल्पनात्मक रूप से ऐसी सरल युक्ति है जो ईएसटी (अभिव्यक्त सीक्वेंस टैग), प्रोटीनों तथा जीनोम इनपुटों के अन्य जीनोमों के बीच जीन सीक्वेंसों की समानता का पता लगाने पर निर्भर है। यह युक्ति उस अवधारणा पर निर्भर है कि कार्यात्मक क्षेत्र (एक्सॉन), अ–कार्यात्मक क्षेत्रों (इंटरजेनिक या इंट्रॉनिक क्षेत्रों) की तुलना में अधिक संरक्षित विकास वाले होते हैं। एक बार जब कुछ जीनोमी क्षेत्रों और किसी ईएसटी, डी.एन.ए. या प्रोटीन के बीच समानता स्थापित हो जाती है तो समानता संबंधी इस सूचना का उपयोग जीन संरचना या उस क्षेत्र के कार्य की व्याख्या करने में किया जा सकता है।

ईएसटी आधारित सीक्वेंस समानता में सामान्यतः कमी यह होती है कि केवल जीन सीक्वेंस के छोटे भागों से सम्बद्ध होते हैं जिसका अर्थ यह है कि किसी निर्धारित क्षेत्र की सम्पूर्ण जीन संरचना का पूर्वानुमान अक्सर कठिन हो जाता है। स्थानीय एलाइनमेंट तथा वैश्विक एलाइनमेंट समानता खोजों पर आधारित दो विधियां हैं। सर्वाधिक सामान्य स्थानीय एलाइनमेंट युक्ति कार्यक्रमों के ब्लास्ट कुल की है जिससे ज्ञात जीनों, प्रोटीनों या ईएसटी के सीक्वेंस की समानता का पता लगाया जाता है। इस प्रकार की युक्ति की सबसे बड़ी कमी यह है कि खोजे गए कुल जीनों में से केवल लगभग आधे जीन ही डेटाबेसों में जीनों के प्रति उल्लेखनीय समांगता प्रदर्शित करते हैं।

एबी इनिशियो जीन पूर्वानुमान विधियां

जीनों की कम्प्यूटेशनल पहचान की विधियों का दूसरा वर्ग वह है कि जिसमें जीनों का पता लगाने के लिए टैम्प्लेट के रूप में जीन संरचना का उपयोग किया जाता है। इसे *एब इनिशियो* पूर्वानुमान भी कहते हैं। *एबी इनिशियो* पूर्वानुमान दो प्रकार की सीक्वेंस सूचना पर निर्भर करते हैं : संकेत सैंसर तथा विषय–वस्तु सैंसर। संकेत सैंसरों का संबंध छोटे सीक्वेंस मॉटिफ से है जैसे स्प्लाइस स्थल, शाखा बिंदु, पॉली पाइरीमिडीन ट्रैक्ट, स्टार्ट कोडॉन तथा स्टॉप कोडॉन। एक्सॉन को ज्ञात करना मुख्यतः विषय–वस्तु सैंसरों पर निर्भर है जिसका अर्थ कोडॉन उपयोग के उन पैटर्नों से है जो प्रजाति विशिष्ट होते हैं तथा जिनसे कोडिंग सीक्वेंसों को सांख्यिकीय पहचान एल्गोरिथ्म द्वारा उनके आस–पास के अ–कोडित सीक्वेंसों से विभेदित किया जा सकता है।

जीन संरचना की मॉडलिंग के लिए अनेक एल्गोरिथ्म उपयोग में लाए जाते हैं जैसे डायनेमिक प्रोग्रामिंग, लीनियर डिस्क्रिमिनेट एनालिसिस, लिंगिस्ट विधियां, हिडन मार्कोब मॉडल तथा न्यूरल नेटवर्क। इन मॉडलों के आधार पर बड़ी संख्या में *एबी इनिशियो* जीन पूर्वानुमान कार्यक्रम विकसित किए गए हैं।

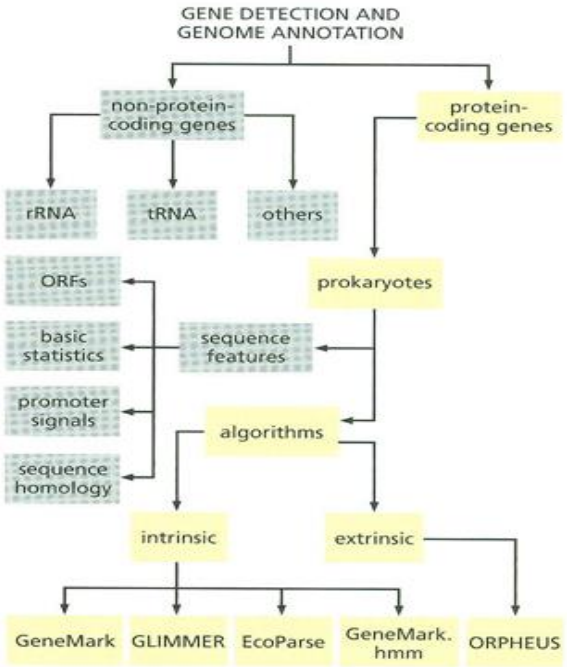


चित्र 2: जीन पूर्वानुमान तथा एनोटेशन का चित्रात्मक प्रस्तुतीकरण

प्रोकैरियोटी जीनोमों में जीन की खोज

प्रोकैरियोटों में विशेष रूप से उच्चतर जीन घनत्व तथा उनके प्रोटीन कोडिंग क्षेत्रों में इंट्रॉनों के न होने के कारण जीनों की खोज अपेक्षाकृत सरल है। वे डी.एन.ए. क्रम जो प्रोटीनों को इनकोड करते हैं, उत्छ। में ट्रांसक्राइब हो जाते हैं तथा उत्छ। सामान्यतः बिना किसी उल्लेखनीय सुधार के प्रोटीनों में ट्रांसलेट हो जाते हैं। उत्छ। के सर्वप्रथम उपलब्ध स्टार्ट कोडॉन से आरंभ होकर समान रीडिंग फ्रेम में अगले स्टॉप कोडोन की ओर जाने वाले सबसे लंबे ओआरएफ (ओपेन रीडिंग फ्रेम) से सामान्यतः प्रोटीन कोडिंग क्षेत्रों का श्रेष्ठ पूर्वानुमान

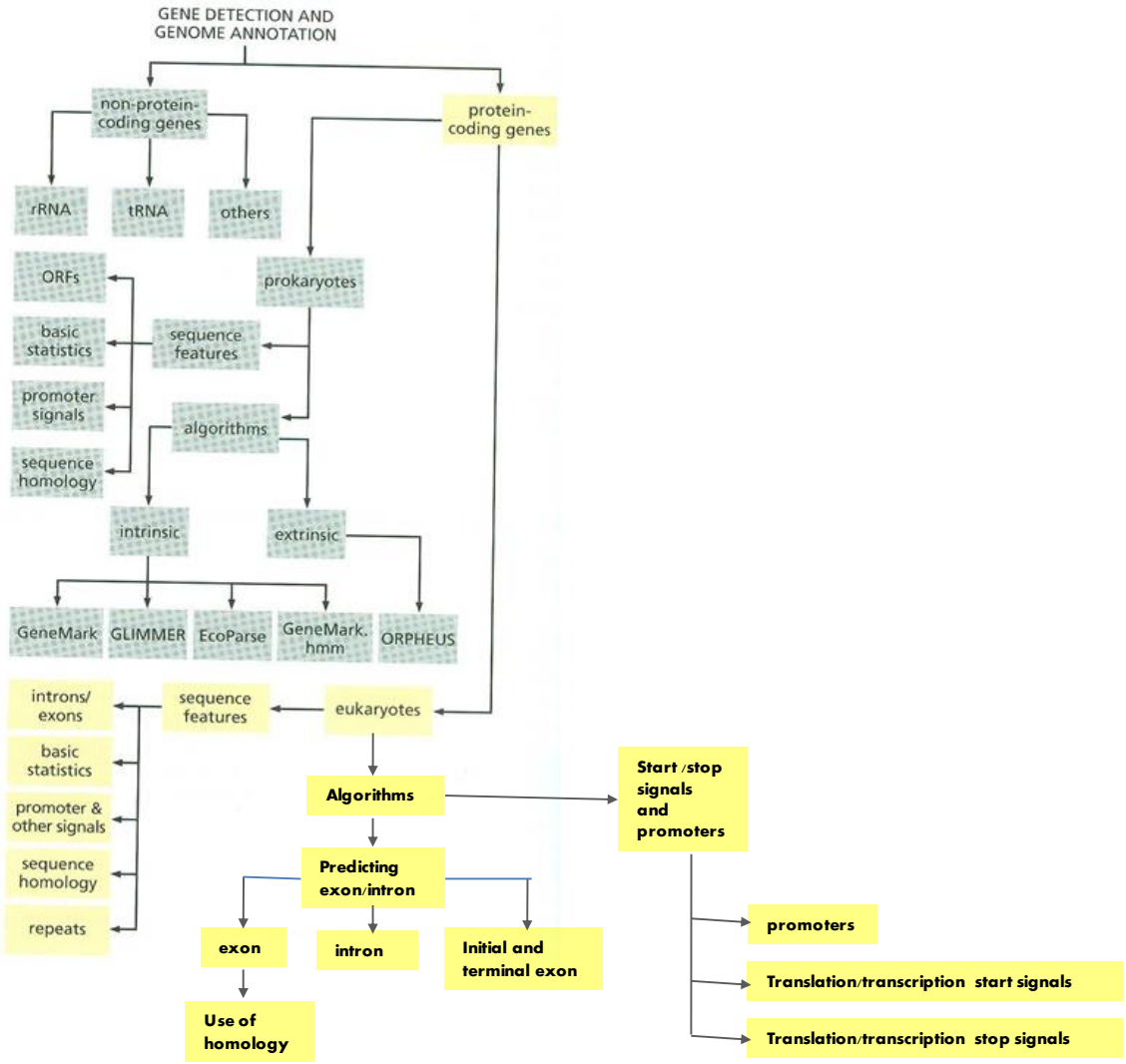
लगता है लेकिन यह पूर्वानुमान सटीक हो, यह सुनिश्चित नहीं होता है। ऐसी अनेक विधियां हैं जिनमें विभिन्न कोडिंग क्षेत्रों जैसे 'छाया' कोडिंग क्षेत्रों (विपरीत डी.एन.ए. लड़ी पर कोडिंग) तथा गैर–कोडिंग डी.एन.ए. के बीच संघटनात्मक भेदों को ज्ञात करना आसान हो जाता है। ऐसी विधियों में इकोपार्स, व्यापक रूप से प्रयुक्त होने वाला जेनमार्क और ग्लीमर कार्यक्रम शामिल हैं जिनसे अच्छे निष्पादन के साथ अधिकांश प्रोटीन कोडिंग जीनों की पहचान की जा सकती है।



चित्र 3: प्रोकैरियोटी जीन खोज का प्रवाह चित्र

यूकैरियोटी जीनोम में जीन की खोज

यह प्रोकैरियाटों में पाई जाने वाली समस्या से बिल्कुल भिन्न समस्या है। विशिष्ट प्रमोटर सीक्वेंसों पर आरंभ किए गए प्रोटीन कोडिंग क्षेत्रों के ट्रांसक्रिप्शन के पश्चात् स्प्लाइसिंग यांत्रिकी द्वारा पूर्व– उत्छ। से नॉन कोडिंग सीक्वेंसों (इंट्रॉनों) को हटाया जाता है जिससे केवल प्रोटीन इनकोडिंग एक्सॉन रह जाते हैं। एक बार जब इंट्रॉन हट जाते हैं तथा परिपक्व आरएनए में कुछ अन्य सुधार कर दिए जाते हैं तो इसके परिणामस्वरूप परिपक्व उत्छ। को 5′ से 3′ दिशा में ट्रांसलेट किया जा सकता है जो सामान्यतः प्रथम स्टार्ट कोडॉन से प्रथम स्टॉप कोडॉन की ओर होता है। यूकैरियाटों के जीनोमी डी.एन.ए. क्रमों में इंट्रॉन सीक्वेंसों की उपस्थिति के परिणामस्वरूप इनकोडित जीन से सम्बद्ध ओआरएफ सामान्यतः इंट्रॉनों की उपस्थिति से बाधित होते हैं जिससे स्टॉप कोडोन सृजित होते हैं (चित्र 4)।



चित्र 4: यूकैरियोटी जीन खोज का प्रवाह चित्र

जीन पूर्वानुमान कार्यक्रम

जीन पूर्वानुमान में दो मूल समस्याएं हैं : प्रोटीन कोडित क्षेत्रों का पूर्वानुमान तथा जीनों के कार्यात्मक स्थलों का पूर्वानुमान। जीन पूर्वानुमान कार्यक्रम को चार पीढ़ियों में वर्गीकृत किया जा सकता है। कार्यक्रमों की प्रथम पीढ़ी को जीनोमी डी.एन.ए. के कोडिंग क्षेत्रों की लगभग स्थितियों का पता लगाने के लिए डिजाइन किया गया है। सर्वाधिक व्यापक रूप से ज्ञात कार्यक्रम संभवतः टेस्ट कोड तथा ळत।प्स् हैं। तथापि, इनसे एक्सॉन की स्थितियों का सटीक पूर्वानुमान नहीं लगाया जा सकता। द्वितीय पीढ़ी में जैसे ैव्त्थ्च्छक् तथा ग्चवनदक स्प्लाइस संकेत तथा कोडिंग क्षेत्र की पहचान की जा सकती है लेकिन इसमें पूर्वानुमानित एक्सॉनों को सम्पूर्ण जीनोमों में असेम्बल करने का प्रयास नहीं किया गया था। कार्यक्रमों की अगली पीढ़ी में जीन की सम्पूर्ण संरचनाओं के पूर्वानुमान जैसे और अधिक कठिन कार्य को सम्पन्न करने का प्रयास किया गया। अनेक कार्यक्रम विकसित किए गए हैं जिनमें जीन आईडी, जीन पार्सल, जीन लैंग तथा थ्ळम्छम्भ् शामिल हैं। तथापि इन कार्यक्रमों का निष्पादन अच्छा नहीं रहा है। इसके अतिरिक्त ये सभी कार्यक्रम इस अवधारणा पर आधारित थे कि इनपुट सीक्वेंस

में ठीक एक सम्पूर्ण जीन होता है जबकि अक्सर ऐसा नहीं होता है। इस समस्या को हल करने तथा सटीकता व उपयोगशीलता को और सुधारने के लिए GENSCAN व AUGUSTUS विकसित किए गए जिन्हें चौथी पीढ़ी के रूप में वर्गीकृत किया जा सकता है।

जीन मार्क

जीन मार्क कोडिंग तथा नॉन–कोडिंग फ्रेमों की सांख्यिकी के प्रस्तुतीकरण के लिए मार्कोव चेन मॉडल का उपयोग करता है। इस विधि में कोडिंग क्षेत्रों की पहचान के लिए डाइकोडोन सांख्यिकी का उपयोग किया जाता है। मान लीजिए कि सीक्वेंस ग जिसकी आधार पर पजी स्थिति x_i होती है। प्रयुक्त मार्कोव चेन पांचवें क्रम की हैं तथा इनमें P(a/x1x2x3x4x5) जो सीक्वेंस ग के छठे आधार की संभाव्यता का प्रतिनिधित्व करता है तथा सीक्वेंस x में जहां x1x2x3x4x5 होते हैं, पूर्व पांच आधारों में व्यक्त किया जाता है जिसके परिणामस्वरूप सीक्वेंस का प्रथम डाइकोडोन x1x2x3x4x5a होता है। ये पद सामान्य सीक्वेंस $b_1b_2b_3b_4b_5$ से युक्त सभी संभावित प्राचलों के साथ परिभाषित किए गए हैं। इन पदों के मान आंकड़ों के विश्वलेषण से प्राप्त किए जा सकते हैं जिसमें वह न्यूक्लियोटाइड सीक्वेंस होता है जिसमें वास्तव में पहचाने गए कोडिंग क्षेत्र होते हैं। जब आंकड़े पर्याप्त होते हैं तो उन्हें निम्न प्रकार से व्यक्त किया जा सकता है :

जहां, ट्रेनिंग डाटा में सीक्वेंस b में व्यक्त संख्या b है। यह ट्रेनिंग डाटा से एस्टीमेटरों की संभाव्यता की सर्वाधिक संभावना है।

ग्लिमर

ग्लिमर का मुख्य तत्व इंटरपोलेटिड मार्कोव मॉडल (आईएमएम) है जिसे प्रसरणशील क्रम के साथ एक सामान्य मार्कोव चेन के रूप में वर्णित किया जा सकता है। जीन मार्क के निर्धारित क्रम की मार्कोव चेन में लागू होने के बाद ग्लिमर जीनोम अंश की मॉडलिंग के लिए बेहतर युक्ति का पता लगाने का प्रयास करता है। प्रोत्साहनपूर्ण तथ्य यह है कि मार्कोव चेन का क्रम जितना बड़ा होगा, नॉन रेंडमनेस को उतने ही अच्छे तरीके से बताया जा सकेगा। तथापि, चूंकि हमें उच्चतर क्रम के मॉडलों की ओर बढ़ना होता है, अतः हमें संभाव्यताओं की उस संख्या का पता लगाना होता है जिससे आंकड़ों को चर घातांकी रूप से बढ़ाया जा सके। स्थिर–क्रम के मार्कोव चेन की मुख्य कमी यह है कि उच्च क्रम के मॉडलों के लिए और अधिक चर घातांकी ट्रेनिंग डाटा की आवश्यकता होती है जो फिलहाल सीमित हैं तथा नए क्रमों में सामान्यतः उपलब्ध नहीं हैं। तथापि, उच्च क्रम से कुछ ऐसे ओलिगोमैर प्राप्त किए गए हैं जो अत्यधिक उपयोगी पूर्वानुमान लगाने के लिए पर्याप्त हैं। इन उच्च स्तर की सांख्यिकियों का उपयोग करने के लिए, जब कभी भी पर्याप्त आंकड़े उपलब्ध होते हैं, ग्लिमर या आईएमएम का उपयोग किया जाता है।

ग्लिमर से 0 क्रम से लेकर 8वें क्रम तक सभी मार्कोव चेनों की संभाव्यताओं की गणना की जा सकती है। यदि लंबे सीक्वेंस (जैसे 8–मैर्स) जल्दी–जल्दी घटित होते हैं तो आईएमएम में आठवें क्रम के मॉडल को ट्रेंड करने के लिए अपर्याप्त आंकड़ों का भी उपयोग किया जाता है। इसी प्रकार जब आठवें क्रम के मॉडल से सांख्यिकी संबंधी उल्लेखनीय सूचना उपलब्ध नहीं होती है तो ग्लिमर में जीनों का पूर्वानुमान लगाने के लिए निम्न–क्रम के मॉडलों का उपयोग किया जाता है।

सुपरवाइज्ड जीन मार्क के विपरीत ग्लिमर में ट्रेनिंग के लिए इनपुट सीक्वेंस का उपयोग होता है। कुछ थ्रेशहोल्ड से लंबे ओआरएफ पहचाने जाते हैं तथा उनका उपयोग ट्रेनिंग के लिए किया जाता है क्योंकि इस बात की अत्यधिक संभावना होती है कि ये प्रोकैरियोटों में पाए जाने वाले जीन हैं। ट्रेनिंग का एक अन्य विकल्प अन्य जीवों से ज्ञात उन जीनों में समांगतता से युक्त क्रमों का उपयोग करना है जो सार्वजनिक डेटाबेसों पर उपलब्ध हैं। तथापि, उपयोगकर्ता यह निर्णय ले सकता है कि ट्रेनिंग के उद्देश्य से लंबे ओआरएफ का उपयोग किया जाए या आईएमएम को ट्रेन करने व निर्मित करने के लिए जीनों के किसी भी सैट को चुना जाए।

जीनमार्क.hmm

जीनमार्क.hmm को सटीक जीन स्टार्टों का पता लगाने के लिए जीनमार्क में सुधार के लिए डिजाइन किया गया है। इसलिए जीनमार्क.hmm के गुण जीनमार्क के सम्पूरक होते हैं। जीनमार्क.hmm में कोडिंग तथा नॉन–कोडिंग क्षेत्रों के जीनमार्क मॉडलों का उपयोग होता है तथा उन्हें हिडेन मार्कोव मॉडल फ्रेमवर्क में शामिल कर लिया जाता है। संक्षेप में कहें तो हिडेन मार्कोव मॉडलों (HMM) का उपयोग नॉन–कोडिंग से कोडिंग क्षेत्रों में ट्रांजिशनों का वर्णन करने या कोडिंग से नॉन कोडिंग क्षेत्रों का वर्णन करने के लिए किया जाता है। जीनमार्क.hmm से वाइटर्बी एल्गोरिथ्म का उपयोग करके जीनोम की संरचना का पूर्वानुमान लगाने की सर्वाधिक संभावना रहती है। यह एल्गोरिथ्म छिपी हुई अवस्थाओं वाले सर्वाधिक संभावित सीक्वेंस का पता लगाने के लिए प्रयुक्त होने वाला एल्गोरिथ्म है। ट्रांसलेशन स्टार्ट स्थिति के पूर्वानुमान में और सुधार करने के लिए जीनमार्क.hmm राइबोसोम बंधन स्थल का एक मॉडल व्युत्पन्न कराता है (स्टार्ट कोडोन के पूर्ववर्ती 6–7 न्यूक्लियोटाइड जो प्रोटीन का ट्रांसलेशन आरंभ होने पर राइबोसोम द्वारा बंधित होते हैं)। इस मॉडल का उपयोग परिणामों को और अधिक सटीक बनाने के लिए किया जाता है।

जीनमार्क और जीनमार्क.hmm से उन ओपेन रीडिंग फ्रेमों के रूप में प्रोकैरियोटी जीनों की पहचान की जाती है जिनमें वास्तविक जीन होते हैं। इसके अतिरिक्त इन दोनों में ट्रेनिंग आंकड़ों के रूप में पूर्व–कम्प्यूटित प्रजाति–विशिष्ट जीन मॉडलों का उपयोग होता है, ताकि प्रोटीन कोडिंग तथा नान–कोडिंग क्षेत्रों के प्राचलों का पता लगाया जा सके।

ओर्फेस

ओर्फेस कार्यक्रम में पूर्व कार्यक्रमों में जिस सूचना की उपेक्षा की गई होती है उसका उपयोग करके विधियों को सुधारने हेतु समांगता, कोडोन संबंधी सांख्यिकी व राइबोसोम बंधन स्थलों का उपयोग किया जाता है। इस विधि में मुख्य भेद यह है कि इसमें प्यूटेटिव जीनों का पता लगाने में डेटाबेस खोजों का उपयोग होता है और इस प्रकार यह एक एक्सट्रिंसिक विधि है। जीनों के इस आरंभिक सैट का उपयोग कोडॉनों के स्तर पर न कि डाइकोडॉनों के स्तर पर कार्य करने के मामले में जीवों की कोडिंग सांख्यिकी को परिभाषित करने के लिए किया जाता है। इसके पश्चात् इस सांख्यिकी का उपयोग प्रत्याशी ओआरएफ के बड़े सैट को परिभाषित करने के लिए किया जाता है। इस सैट से स्पष्ट स्टार्ट कोडॉन छोर से युक्त ओआरएफ का उपयोग राइबोसोम बंधन स्थल के लिए स्कोर मैट्रिक्स को परिभाषित करने के लिए किया जाता है जिसे आगे चलकर उन ओआरएफ के 5′ छोर को निर्धारित करने के लिए किया जाता है जहां वैकल्पिक स्टार्ट मौजूद होता है।

इकोपार्स

जीन फाइंडर पर आधारित प्रथम एचएमएम मॉडल इकोपार्स को *ई.कोलाई* में जीन का पता लगाने के लिए विकसित किया गया था। इसमें जीनों के कोडोन की संरचना के उपयोगों पर ध्यान केन्द्रित किया जाता है। एचएमएम आधारित जीन फाइंडर के फलोरा युक्त ईकोपार्स एक गतिशील कार्यक्रम है तथा यह सीक्वेंस को पार्स करने के लिए एक वाइटेब्री एल्गोरिथ्म के रूप में उभर कर सामने आया है।

जीनपूर्वानुमान कार्यक्रमों का विकास

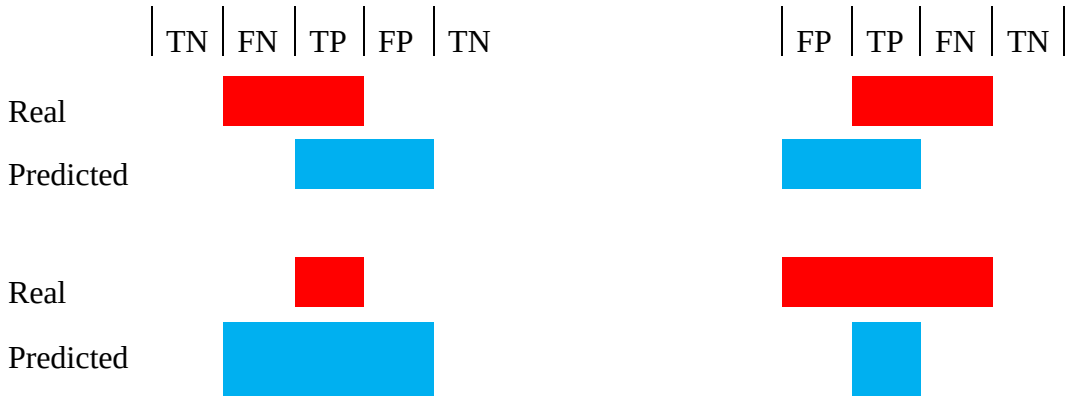
जीनपूर्वानुमान के क्षेत्र में सटीकता को तीन स्तरों पर मापा जाता है :

क. कोड किए गए न्यूक्लियोटाइड (आधार स्तर)

ख. एक्सॉन संरचना (एक्सॉन स्तर)

ग. प्रोटीन उत्पादन (प्रोटीन स्तर)

आधार स्तर के जीन पूर्वानुमानों को *ट्रू पोजीटिव्स (टीपी)* (वे पूर्वानुमानित गुण जो वास्तविक होते हैं), *ट्रू नेगेटिव* (टीएन) (वे अपूर्वानुमानित गुण जो वास्तविक नहीं होते हैं), *फाल्स पोजिटिव* (एफटी) (वे पूर्वानुमानित गुण जो वास्तविक नहीं होते हैं) तथा *फाल्स नेगेटिव* (एफएम) (वास्तविक गुण जो पूर्वानुमानित नहीं थे) चित्र 5। सामान्यतः आधार एलाइनमेंट कोडिंग या नॉनकोडिंग खण्ड में होना चाहिए लेकिन इस विश्लेषण को जीनों के नॉन कोडिंग भागों या सीक्वेंसों के किसी भी कार्यात्मक भाग को शामिल करने के लिए और बढ़ाया जा सकता है।



चित्र 5. वास्तविक और पूर्वानुमानित जीनां की चार संभावित तुलनाएं

संवेदनशीलता (Sn) : वास्तविक जीनों में उन आधारों के अंश जिन्हें जीनों में सटीक रूप से पूर्वानुमानित किया जाना होता है, संवेदनशीलता है तथा इसे किसी दिए हुए जीन में न्यूक्लियोटाइड को सटीक रूप से पूर्वानुमानित करने के लिए संभाव्यता के रूप में निम्नानुसार व्याखित किया जाता है।

विशिष्टता (Sp) : उन आधारों का अंश जिन्हें वास्तव में मौजूद जीनों के पूर्वानुमान के लिए प्रयुक्त किया जाता है, विशिष्टता कहलाते हैं तथा इन्हें किसी दिए गए जीन में वास्तव में उपस्थित न्यूक्लियोटाइड की संभावना के रूप में निम्नानुसार व्याखित किया जा सकता है।

जीन पूर्वानुमान कार्यक्रम के इन दो मानों के उपयोग में सावधानी बरती जानी चाहिए क्योंकि विशिष्टता की सामान्य परिभाषा के अनुसार अत्यधिक उच्च परिणाम गलत भी हो सकते हैं।

इन कठिनाइयों को दूर करने के एकल उपाय के रूप में एप्रोक्सीमेट कोरिलेशन कोफिसिएंट (AC) प्रस्तावित किया गया है। इसे AC=2 (ACP-0.5) के रूप में परिभाषित किया गया है जहां

एक्सॉन के स्तर पर पूर्वानुमान निर्धारण की सटीकता एक्सॉन स्टार्ट तथा एंड प्वाइंट्स के सटीक पूर्वानुमान पर निर्भर करती है। इस क्षेत्र में प्रयुक्त होने वाली संवेदनशीलता तथा विशिष्टता के दो प्रमुख माप हैं, इनमें से प्रत्येक माप यद्यपि भिन्न है लकिन यह अत्यधिक उपयोगी है।

संवेदनशीलता की मापों में निम्न का उपयोग होता है :

S_{n1} = CE/AE और Sn2 =ME/AE

विशिष्टता की मापों में निम्न का उपयोग होता है :

S_{p1} = CE/PE और Sp2 =WE/PE

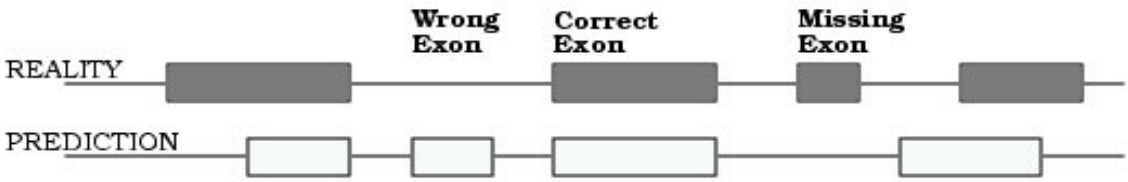
यहां AE= आंकड़ों में वास्तविक एक्सॉन की संख्या

PE= आंकड़ों में पूर्वानुमानित एक्सॉनों की संख्या

CE= सही पूर्वानुमानित एक्सॉनों की संख्या

ME= गायब एक्सानों की संख्या (यदा–कदा ही होता है)

WE= गलत पूर्वानुमानित एक्सॉनों की संख्या (चित्र 5)



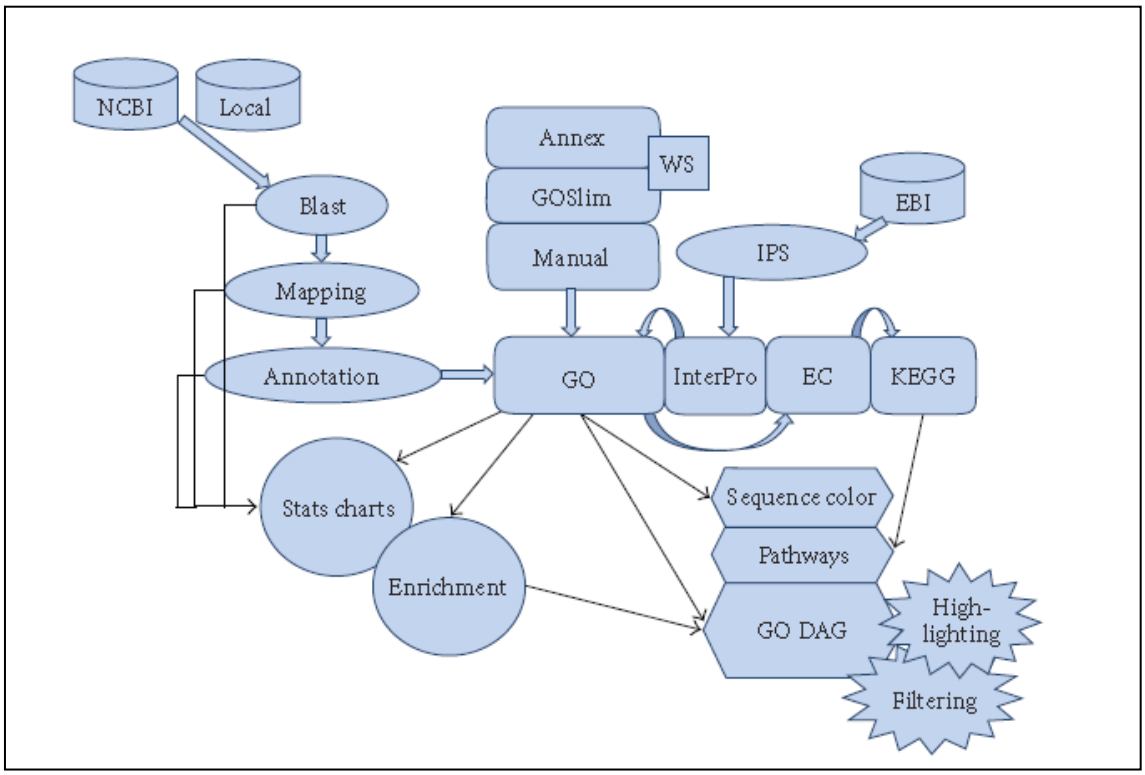
चित्र 6. वास्तविक तथा पूर्वानुमनित एक्सॉन

जीन आंटोलॉजी

जीन आंटोलॉजी (GO, http:www/geneotology.org) संभवतः जीन उत्पाद कार्यों के वर्णन के लिए वर्तमान में उपलब्ध सबसे गहन स्कीम है लेकिन अन्य प्रणालियां जैसे एंजाइम कोड, केईजीजी पथ, फनकैट या सीओजी का भी व्यापक रूप से उपयोग हो रहा है। यहां हम कार्यात्मक एनोटेशन, प्रबंध तथा नए सीक्वेंस की डेटा माइनिंग के लिए ब्लास्ट2GO (B2G, www.blast2go.org) एप्लिकेशन का वर्णन कर रहे हैं जिसमें सामान्य नियंत्रित शब्दावली स्कीम का उपयोग होता है। इस युक्ति का सामान्य अनुप्रयोग क्षेत्र नॉन मॉडल जीनों में कार्यात्मक जीनोमिक्स है तथा इसका उद्देश्य प्राथमिकतः प्रयोगात्मक प्रयोगशालाओं में अनुसंधान कार्य में सहायता पहुंचाना है। blast2GO में कार्यात्मक जीनोमिक्स परियोजनाओं के अंतर्गत नए सीक्वेंसों के एनोटेशन के लिए अनुप्रयोग की विधि का चुनाव किया जाता है

जहां हजारों खण्डों का लक्षण–वर्णन करने की आवश्यकता होती है। इसेंज2ळ८ में कार्यात्मक एनोटेशन समांगता हस्तांतरण पर आधारित है। इस फ्रेम वर्क में वास्तविकएनोटेशन क्रियाविधि कन्फीगुरेट की जा सकती है तथा इसमें विभिन्न एनोटेशन कार्यनीतियों के डिजाइन के उपयोग की अनुमति है। blast2GO एनोटेशन प्राचलों में खोज डेटोबेस की पसंद, ब्लास्ट परिणामों की शक्ति तथा संख्या, क्वैरी–हिट मैच की सीमा, हस्तांतरित एनोटेशनों की गुणवत्ता तथा मॉटिफ एनोटेशन को शामिल करना जैसे पहलू शामिल हैं। ठ2ळ द्वारा समर्थित शब्दावलियां जीन ओंटोलॉजी पदावली, एंजाइम कोड (म्ब), इंटरप्रो आईडी तथा ज्ञम्ळळ पाथवे हैं।

चित्र 7 में blast2GO स्यूट के मूल घटक दर्शाए गए हैं। कार्यात्मक निर्धारण ऐसी विस्तृत एनोटेशन क्रियाविधि के द्वारा किया जाता है जिसमें केन्द्रीय कार्यनीति के साथ–साथ कार्यों का परिशोधन भी शामिल है। इसके बाद विजियोलाइजेशन तथा डेटा माइनिंग इंजनों से कार्यात्मक ज्ञान प्राप्त करने के लिए एनोटेशन परिणामों का उपयोग किया जाता है। GO एनोटेशन 3–चरण वाली प्रक्रिया से सृजित होते हैं : ब्लास्ट, मैपिंग, एनोटेशन। इंटर प्रो शब्द ईबीआई पर इंटरप्रो स्कैन से प्राप्त होते हैं तथा उन्हें GOs में परिवर्तित करके समाहित किया जाता है। GO एनोटेशन को ।ददमगए GOSlim वेब सेवाओं तथा मानवीय संपादन द्वारा मॉड्यूलेट किया जा सकता है। EC तथा KEGG एनोटेशन GO से सृजित होते हैं। दृष्टव्य युक्तियों में सीक्वेंस रंग कोड, KEGG पथ तथा GO ग्राफ शामिल हैं जिनमें नोड हाइलाइटिंग तथा फिल्टरिंग विकल्प होते हैं। अतिरिक्त एनोटेशन डेटा–माइनिंग युक्तियों में सांख्यिकी चार्ट तथा जीन सैट समृद्धिकरण संबंधी विश्लेषण कार्य शामिल हैं।



चित्र 7: Blast2GO **एप्लिकेशन का रेखाचित्र द्वारा प्रस्तुतीकरण**

Blast2GO एनोटेशन प्रक्रिया में तीन मुख्य चरण हैं : समांगी सीक्वेंसों का पता लगाने के लिए ब्लास्ट, ब्लास्ट हिटों से संबंधित ळ्व पदों को एकत्रित करने के लिए मानचित्रण तथा क्वेरी सीक्वेंसों को भरोसेमंद सूचना प्रदान करने के लिए एनोटेशन।

ब्लास्ट चरण

B2G में प्रथम चरण ब्लास्ट द्वारा सैट की गई क्वैरी के समान सीक्वेंसों का पता लगाना है। B2G FASTA फार्मेट में न्यूक्लियोटाइड तथा प्रोटीन सीक्वेंसों को स्वीकार करता है तथा मूल ब्लास्ट कार्यक्रमों (चार–ब्लास्टग, ब्लास्टचए ब्लास्टद और ब्लास्टग) को सहायता पहुंचाता है। समांगता संबंधी खोजें सार्वजनिक डेटाबेसों से आरंभ की जा सकती हैं जैसे ब्लास्ट के क्वेरी फ्रेंडली वर्जन (QBlast) का उपयोग करके NCBI दत। यह एक डिफाल्ट विकल्प है और इस मामले में किसी अतिरिक्त इंस्टालेशन की आवश्यकता नहीं होती है। विकल्प के रूप में ब्लास्ट को उचित रूप से FASTA–फार्मेट किए गए डेटाबेस के विरुद्ध स्थानीय रूप से चलाया जा सकता है जिसके लिए कार्यशील www.blast की आवश्यकता होती है। टूल्स मेन्यू में मेक फिल्टर ब्लास्ट–ळ्ठक कार्य से केवल ळ्व एनोटिड प्रविष्टियों से युक्त डेटाबेसों को कस्टमाइज्ड करते हुए सृजित किया जा सकता है जिसका उपयोग स्थानीय ब्लास्ट विकल्प के साथ मिलाकर किया जा सकता है। ब्लास्ट चरण पर एक अन्य कंफीगुरेशन युक्त प्राचल अपेक्षा मान (ई–मान) थ्रेशहोल्ड है जिसमें प्राप्त किए गए हिटों की संख्या तथा न्यूनतम एलाइनमेंट लंबाई (एचएसपी लंबाई) ज्ञात होती है जिससे छोटे, निम्न ई–मानों से मेल खाते हिटों का निष्कासन कार्यात्मक पदों के स्रोतों से करना संभव होता है। तथापि, एनोटेशन अंततः क्रमों की समानता के स्तरों पर आधारित होता है क्योंकि सामान्यता के प्रतिशत डेटाबेस के आकार से स्वतंत्र होते हैं तथा ई–मानों की तुलना में अधिक इंट्यूटिव होते हैं। Blast2GO ब्लास्ट परिणामों को पार्स करता है तथा प्रत्येक सीक्वेंस के लिए तालिका स्वरूप में सूचना को प्रस्तुत करता है। क्वैरी सूचना संबंधी विवरण हिट विवरणों में भाषा प्रसंस्करण एल्गोरिथ्म का उपयोग करके प्राप्त किए जाते हैं जो सूचनात्मक नामों से प्राप्त होते हैं तथा इनमें निम्न अंश वाले पद जैसे 'हाइपोथैटिकल प्रोटीन' या 'एक्सप्रेस्ड प्रोटीन' नहीं होते हैं।

मानचित्रण चरण

मानचित्रण जीओ पदों से संबंधित सूचना को प्राप्त करने की प्रक्रिया है ताकि ब्लास्ट खोज के पश्चात् कितने हिट हुए हैं, इसे ज्ञात किया जा सके। B2G में तीन भिन्न मानचित्रण निष्पादित होते हैं, जो निम्नानुसार हैं :

क. ब्लास्ट परिणाम संबंधी प्रविष्टियों का उपयोग एनसीबीआई द्वारा उपलब्ध कराई गई दो मानचित्रण फाइलों (geneinfo, gene2accession) का उपयोग करके जीनों के नामों (संकेतों) को प्राप्त करने के लिए भी किया जाता है। पहचाने गए जीन नामों को GO2 डेटाबेस की जीन उत्पाद तालिका की प्रजाति–विशिष्ट प्रविष्टियों में खोजा जाता है।

ख. ब्लास्ट परिणाम जीआई आइडेंटीफायर्स का उपयोग पीएसडी, यूनिपोर्ट, स्वीस–पोर्ट, ट्रीएम्बल, रैफसैक, जैनपैप्ट तथा पीडीबी सहित पीआईआर (नॉन–रिडंडेंट रैफरेंस प्रोटीन डेटाबेस) से मानचित्रण फाइलों का उपयोग करने के लिए यूनिपोर्ट आईडीस को प्राप्त करने के लिए किया जाता है।

ग. ब्लास्ट परिणाम संबंधी प्रविष्टियों को GO डेटाबेस की DBXRef तालिका में सीधे खोजा जाता है।

एनोटेशन चरण

यह मानचित्रण चरण में एकत्र किए गए ळव् पदों के पूल से क्वेरी सीक्वेंसों को कार्यात्मक पद देने की प्रक्रिया है। कार्यात्मक एसाइनमेंट जीन आंटोलॉजी की शब्दावली पर आधारित है। GO पदों से एंजाइम कोडों के मानचित्रण से एंजाइम कोडों तथा KEGG पथ एनोटेशनों को अंततः प्राप्त किया जा सकता है। B2G एनोटेशन एल्गोरिथ्म में क्वैरी तथा हिट सीक्वेंसों के बीच समानता, GO एसाइनमेंट स्रोतों की गुणवत्ता तथा ळव् क।ळ की संरचना को ध्यान में रखा जाता है। प्रत्येक क्वेरी सीक्वेंस तथा प्रत्येक प्रत्याशी ळव् पद के लिए एनोटेशन स्कोर (एएस) की गणना की जाती है (चित्र 8 कृपया देखें)। एएस दो पदों से बना है, पहला प्रत्यक्ष पद (डीटी) है जो इस GO पद में मौजूद हिट सीक्वेंसों के बीच सर्वोच्च समानता के मान को दर्शाता है जिसे इसके प्रमाण कोड (EC) के सम्बद्ध घटक के रूप में मापा जाता है। GO का एक पद ईसी है जो कार्यात्मक एसाइनमेंट की प्रक्रिया को इंगित करने के लिए GO डेटाबेस में प्रत्येक एनोटेशन के लिए उपस्थित होता है।

$$DT = \max(\text{similarity} \times EC_{\text{weight}})$$
$$AT = (\#GO - 1) \times GO_{\text{weight}}$$
$$AR : \text{lowest.node}(AS(DT + AT) \geq \text{threshold})$$

चित्र 8: Blast2GO **एनोटेशन नियम**

ईसी प्रायोगिक परिणामों से भिन्न होते हैं जैसे अनसुपरवाइज्ड एसाइनमेंटों से प्रत्यक्ष मूल्यांकित (आईडीए) जिनकी व्याख्या इलेक्ट्रॉनिक एनोटेशन (आईईए) द्वारा की जाती है। दूसरा पद एनोटेशन नियम से संबंधित है (एटी) जिससे एनोटेशन एल्गोरिथ्म में अमूर्त की संभावना व्यक्त होती है। अमूर्तता को पूर्वज नोड के एनोटेशन के रूप में तब परिभाषित किया जाता है जब GO प्रत्याशी पूल में अनेक चाइल्ड नोड मौजूद होते हैं। यह पद उपयोगकर्ता द्वारा परिभाषित घटक या GO वेट (GOw) द्वारा नोड पर एकीकृत कुल GOs की संख्या को प्रगुणित करता है जिससे अमूर्तता की संभाव्यता तथा शक्ति नियंत्रित होती है। जब सभी ईसीडब्ल्यू 1 (कोई ईसी नियंत्रण नहीं) पर सैट होते हैं और GO शून्य (कोई अमूर्तता संभव नहीं) पर सैट होते हैं तो एनोटेशन स्कोर को उस पद से एनोटिड ब्लास्ट हिटों के बीच सर्वोच्च समानता मान के बराबर GO पद के रूप में व्यक्त किया जाता है। यदि ईसीडब्ल्यू एक से कम होता है तो डीटी बढ़ जाता है तथा एनोटेशन थ्रेशहोल्ड को पार करने के लिए उच्चतर क्वेरी–हिट समानताओं की आवश्यकता होती है। यदि GO शून्य के बराबर नहीं होता है तो एटी योगदाता बन जाता है तथा पूर्वज नोड का एनोटेशन संभव हो जाता है बशर्ते कि अनेक चाइल्ड नोड एक साथ विद्यमान हों तथा वे इनोटेशन कट आफ तक न पहुंचते हों। B2G एनोटेशन प्राचलों के डिफाल्ट मान एनोटेशन कवरेज तथा एनोटेशनों की परिशुद्धता के बीच के अनुपात को उपयुक्ततम बनाने के लिए चुने गए थे। एआर प्रति शाखा न्यूनतम पदों को चुनता है जो उपयोगकर्ता द्वारा परिभाषित थ्रेशहोल्ड के आगे बढ़ जाते हैं।

Blast2GO में उपरोक्त परिभाषित प्रक्रिया के माध्यम से प्राप्त किए गए एनोटेशनों को पूरा करने व उन्हें सुधारने के लिए विभिन्न कार्यों को शामिल किया जाता है। एंजाइम कोड तथा KEGG पथ एनोटेशन एंजाइम कोड समतुल्यता के लिए GO पदों के प्रत्यक्ष मानचित्रण से सृजित होते हैं। इसके अतिरिक्त Blast2GO में B2G इंटरफेस से सीधे इंटरप्रो खोजें की जा सकती हैं। B2G में बैच में सीक्वेंस क्वेरी लॉच किए जा सकते हैं तथा इससे इंटरप्रो परिणामों

को प्राप्त, पार्स और अपलोड किया जा सकता है। इसके अतिरिक्त इंटरप्रो आईडी को ळ्व् पदों के रूप में मानचित्रित करते हुए ब्लास्ट–व्युत्पन्न GO एनोटेशनों में समाहित किया जा सकता है, ताकि एक समेकित एनोटेशन परिणाम उपलब्ध हो सके। इस प्रक्रिया में B2G से यह सुनिश्चित होता है कि प्रत्येक शाखा में केवल न्यूनतम पद ही अंतिम एनोटेशन सैट में बना रहे तथा समाहन क्रिया से उत्पन्न होने वाले संभावित पैरंट–चाइल्ड संबंधों की सभी संभावनाओं को समाप्त किया जा सके।

संदर्भः

1. कोनेसा, एस. गोट्ज, जे.एम. ग्रासिया–गोमेज, जे. टेरोल, एम. टालन, और एम. रोबेल्स, 'Blast2GOः ए यूनिवर्सल टूल फॉर एनोटेशन विजुलाइजेशन एंड एनालिसिस इन फंक्शनल जीनोमिक्स रिसर्च' बायोइंफोर्मेटिक्स, खण्ड 21, अंक 18, मु.पृ. 3674–3676, 2005.
2. कोनेसा और एस . गोट्ज, 'Blast2GOः ए कम्प्रीहैंसिव स्यूट फॉर फंक्शनल एनालिसिस इन प्लांट जीनोमिक्स', इंटरनेशनल जर्नल ऑफ प्लांट जीनोमिक्स, खण्ड 2008, 2008.
3. एच. ओगाटा, एस. गोटो, के. साटो, डब्ल्यू फ्यूजीबुची, एच. बोनो और एम. कानेहिसा, 'केईजीजी : क्योटो इनसाइक्लोपीडिया ऑफ जीन्स एवं जीनोम्स', न्यूक्लिइक एसिड्स रिसर्च, खण्ड 27, अंक 1, मु.पृ. 29–34, 1999.
4. जे.डी. वाट्सन, आर.एम. मेयर्स, एए क्वाडी और जे.ए. विट्कोवत्स्की, 'रिकोम्बीनेंट डी. एन.ए. : जीन्स एंड जीनोम्स – ए शॉट कोर्स' तृतीय संस्करण, 2007
5. एम. एशबर्नर, सी. ए. बाल, जे.ए. ब्लाक और साथी, 'जीन आंटोलॉजी : टूल फॉर यूनिफिकेशन ऑफ बायोलॉजी. द जीन आंटोलॉजी कंसोर्टियम', नेचर जेनेटिक्स, खण्ड 25, अंक 1, मु.पृ. 25–29, 2000.
6. रूएप, ए. जोलनेर, डी. माइएर और साथी, 'द फन कैट, ए फंक्शनल एनोटेशन स्कीम फॉर सिस्टेमेटिक क्लासीफिकेशन ऑफ प्रोटीन्स फ्राम होल जीनोम्स, 'न्यूक्लिइक एसिड्स रिसर्च' खण्ड 32, अंक 18, मु.पृ. 5539–5545, 2004.
7. आर.एल. तातुसोव, एन.डी. फैदोरोवा, जे.डी. जैक्सर और साथी, 'द सीओजी डेटाबेसः एन अपडेटिड वर्जन इन्क्लुड्स यूकैरियोट्स,' बीएमसी बायोइंफोर्मेटिक्स, खण्ड 4, पृ. 41, 2003.
8. स्कोम्बर्ग, ए. चांग, सी. एबेलिंग और साथी, 'ब्रेंडा, द एंजाइम डेटाबेस : अपडेट्स एंड मेजर न्यू डेवलपमेंट्स,' न्यूक्लिइक एसिड रिसर्च, खड 32, डेटाबेस अंक, मु.पृ.डी 431–डी 433, 2004.
9. एस.एफ. एल्टस्चल, डब्ल्यू. गिश, डब्ल्यू. मिलर, ई.डब्ल्यू मेयर्स और डीजे लिपमैन, 'बेसिक लोकल एलाइनमेंट सर्च टूल', जर्नल ऑफ मॉलीक्यूलर बायोलॉजी, खण्ड 215, अंक 3, मु.पृ. 403–410, 1990.
10. एस. म्याहरे, एच. त्वेइत, टी. मोलेस्ताद, ए. लीग्रेइड, 'एडीशनल जीन आंटोलॉजी स्ट्रक्चर फॉर इम्प्रूव्ड बायोलॉजिकल रीजनिंग', बायोइंफोर्मेटिक्स, खण्ड 22, अंक 16, मु.पृ. 2020– 2027, 2006.

जीनोमिक चयन के लिए सांख्यिकीय तरीके

पारंपरिक आनुवंशिक सुधार हेतु फिनोटाइप और पेडीग्री का उपयोग कर प्रजनन मूल्यों का अनुमान काफी सफल पाया गया है। हालांकि, जानवरों और पौधों में डीएनए अनुक्रम में बदलाव की जानकारी का उपयोग करके प्रजनन मूल्यों की सटीक भविष्यवाणी की जा सकती है। मार्कर असिस्टेड सलेक्शन की दिशा में व्यापक अनुसंधान हुआ है लेकिन इसका कार्यान्वयन अभी भी सीमित ही है। जीनोमिक सलेक्शन इन कमियों को दूर करने के लिए प्रस्तावित किया गया है। जीनोमिक सलेक्शन मार्कर असिस्टेड सलेक्शन की आधुनिक अवस्था है जिसमें आनुवंशिक मार्कर पूरे जीनोम को कवर करने के लिए इस्तेमाल होता है, ताकि सभी क्वानटीटेटीव ट्रेट लोसाइ (क्यू टी एल) कम से कम एक मार्कर के साथ में लिंकेज डिसइक्यूलीव्रीयम (एल डी) में हों। जीनोमिक सलेक्शन फिनोटाइप और उच्च घनत्व मार्कर का विश्लेषण करके पंक्तियों की प्रजनन मूल्यों की भविष्यवाणी करता है। फिनोटाइप के सलेक्शन एवं पूर्वानूमान करने के लिए विभिन्न तकनीकियाँ विकसित की गयी हैं। ये तकनीकियाँ जीनोटाइप और फिनोटाइप की जानकारी के आंकड़ों के विश्लेषण पर आधारित हैं।

कुंजी शब्द: क्वानटीटेटीव ट्रेट लोसाइ (क्यू टी एल), लिंकेज डिसइक्यूलीव्रीयम (एल डी), मार्कर एसिसटेड सलेक्शन, जीनोमिक सलेक्शन, इपिस्टेसिस

परिचय

जैसा की यह ज्ञात है कि फिनोटीपिक डेटा पर आधारित सलेक्शन अतीत में काफी सफलतापूर्वक इस्तेमाल किया गया है। डी एन ए और मार्कर डेटा की बहुतायत होने से, मार्कर असिस्टेड सलेक्शन का प्रयोग काफी बढ़ गया है। मार्कर असिस्टेड सलेक्शन एक अप्रत्यक्ष सलेक्शन प्रक्रिया है जहां लक्षण व्यक्तिगत विशेषता के आधार पर चुना जाता है। मार्कर एसिसटेड सलेक्शन उन जीन के लिये बहुत उपयोगी सिद्ध होता है जब कोई लक्षण बड़े प्रभाव के कुछ प्रमुख जीन के साथ जुड़े रहते हैं, लेकिन जब यह पोलीजेनिक लक्षण के सलेक्शन के लिए प्रयोग किया जाता है तब उतना अच्छा प्रदर्शन नही करता है (बरनार्डो 2009)। लगभग सभी इकोनॉमिक ट्रैट को बहुत सारे जीन प्रभावित करते हैं लेकिन डी एन ए मार्कर द्वारा सिर्फ कुछ जीनों की व्याख्या करने से बहुत कम जेनेटिक वैरिएंस का वर्णन हो पाता है। इसके अलावा, व्यक्तिगत जीन के बहुत छोटा प्रभाव होने की संभावना रहती है अतः इन प्रभाव की सही भविष्यवाणी के लिये डेटा की बड़ी मात्रा की आवश्यकता होती है।

इन कठिनाइयों को दूर करने के लिए, मिउविसीन आदि (2001) ने मार्कर एसिसटेड सलेक्शन का एक अन्य प्रकार दिया, जिसे जीनोमिक सलेक्शन कहा जाता है। इस विधि की प्रमुख विशेषता है कि मार्कर पूरे जीनोम को कवर करने के लिये इस्तेमाल होता है जिससे की मार्कर्स द्वारा सभी जेनेटीक वैरिएसंस का अध्यन एवं वर्णन किया जा सके। जीनोमिक सलेक्शन के कार्यान्वयन की प्रमुख सीमा है कि इसमें बड़ी संख्या में मार्कर्स की आवश्यकता होती है तथा इन मार्कर्स की जीनोटाइपिंग लागत बहुत अधिक है। हाल ही में पशुओं और पौधों की प्रजातियों के जीनोम अनुक्रमण के बाद तथा हजारों सिंगल न्युकिलिओटाइड पॉलिमॉरफिज्म (एस एन पी) की उपलब्धता और एस एन पी जीनोटाइपिंग प्रौद्योगिकी के विकास में व्यापक सुधार सें ये कमियां दूर हो गयी है। फिनोटाइप की भविष्यवाणी के लिए विभिन्न रिग्रेसन तकनीकियों को विकसित किया गया है। ये तकनीकियाँ जीनोटाइप तथा फिनोटाइप आंकड़ों के विश्लेषण पर आधारित है। इन तकनीकियों में मुख्य रूप से रेखीय मॉडल है, जो कि आंकड़ों की व्याख्या बिना ओवरफिटिंग के करने में सक्षम है। हालांकि, प्रजनन मूल्य और आनुवंशिक मार्कर्स के बीच संबन्ध थोड़ा जटिल है, विशेष रूप से जब बड़ी संख्या में एस एन पी की फिटिंग मॉडल में एक साथ करते है तब एक सरल रैखिक संबंध की तुलना में अधिक जटिल संबंध होने की संभावना रहती है। इन प्रश्नो का उत्तर देने के लिये मॉडल-मुक्त या तथाकथित नानपेरामेट्रीक तरीके का प्रयोग करके अधिक ध्यान आकर्षित करते हैं जिसमें कि कम सांख्यिकीय मान्यताओं (एजम्पसंस) की आवश्यकता होती है (जिनोला आदि 2006)। इस अध्ययन में केवल पैरामीट्रिक जीनोमिक सलेक्शन के तरीकों की चर्चा की गयी है।

जीनोमिक सलेक्शन में सबसे अधिक इस्तेमाल होने वाले पैरामीट्रिक तरीकों का अवलोकन:

जीनोमिक सलेक्शन का मुख्य लक्ष्य व्यक्ति की जीनोटाइप और फिनोटाइप के बीच संबंधों को मॉडलिंग करके व्यक्ति की प्रजनन मूल्य की भविष्यवाणी करना है। इस तरह के मॉडल का एक सरलतम रूप है:

$$Y_i = \mu + \sum_{j=1}^{p} X_{ij}\beta_j + e_i$$

जहां μ एक इंटरसेप्ट है, X_{ij} में i^{th} व्यक्ति के जीनोटाइप j^{th} मार्कर पर है, जहाँ $j = 1,2,..., p$ है, और β_j इसी मार्कर का प्रभाव है। यह साधारणत: एक रेखीय मॉडल का रूप लेता है

$$Y = X\beta + e$$

$(\hat{\mu}, \hat{\beta})$ को इस्टीमेट करने के लिए हम लिस्ट स्कवॉयर विधि का उपयोग कर सकते हैं

$$(\hat{\mu}, \hat{\beta}) = (Y - X\beta)'(Y - X\beta)$$

$$\hat{\beta} = (\acute{X}X)^{-1}\acute{X}Y$$

डिजाइन मैट्रिक्स X में प्रविष्टियां अलग-अलग एलिल्स की संख्या पर निर्भर करती है। उदाहरण के लिए, व्यक्ति जिसका मार्कर जीनोटाइप AA, Aa, aa के तत्व X_{ij} में क्रमश: -1, 0, और 1 के रूप में लिया जाता है। उन्नत जीनोटाइपिंग प्रौद्योगिकियों के होने से मार्कर से उत्पन्न डेटा बहुत बड़ा होता है इसलिए यहां प्राप्त मानकों की संख्या टिप्पणियों की संख्या से अधिक हो सकती है। $p > n$ जैसी समस्याओं का सामना करने के लिए वैरिएवल सलेक्शन, श्रीकेज ऑफ इस्टिमेट या दोनों के संयोजन आमतौर पर इस्तेमाल किया जाता है। लेकिन यह तब भी खराब प्रदर्शन कर सकते हैं यदि मार्कर की संख्या व्यक्तियों या अभिलेखों की संख्या के अनुपात में बड़ी होती है। इसलिए, दूसरा पसंदीदा मॉडल संकोचन आकलन (श्रीकेज इस्टीमेसन) की प्रक्रिया का उपयोग करते हैं।

किसी भी आकलनकर्ता की शुद्धता की जाँच अनुमानित $\hat{\beta}$, तथा वास्तविक $\hat{\beta}$, के बीच दूरी की गणना करके की जा सकती है। इसे हम साधारणत: परिभाषित कर सकते है: $\left\|\hat{\beta}(Y)-\beta\right\|^2 = \left[\hat{\beta}(Y)-\beta\right]^2$ तथा मीन स्कवेयर एरर को हम इस प्रकार परिभाषित कर सकते है: $MSE\left(\hat{\beta}\right) = E\left[\hat{\beta}(Y)-\beta\right]^2$

इसके अलावा हम इसे दो भाग में विघटित कर सकते हैं: $MSE\left(\hat{\beta}\right) = Var\left(\hat{\beta}\right) + Bias(\beta)^2$

मानक आकलन की प्रक्रिया में सैम्पल साइज बढने सें आकलनकर्ता का विचरण घटता है, तथा सैम्पल साइज को फिक्स कर p को वढाने से आकलनकर्ता का विचरण काफी बढ जाता है। अनुमान को एक निश्चित बिंदु की ओर स्रीकेज करने से p>n कि समस्या को हल करते हैं। पेनेलाइज्ड, बेसियन और बेसियन लासो पद्धति सबसे व्यापक रूप से इस्तेमाल होने वाली संकोचन प्रक्रिया हैं, इन तरीकों कि संक्षिप्त में नीचे चर्चा की गयी है।

पेनेलाइज्ड तरीके

पेनेलाइज्ड तरीकों में, प्रतिगमन अनुमान की गणना प्रशिक्षण डेटा के मॉडल फिट तथा मॉडल की जटिलता को नियंत्रित करने के लिए करते हैं। मॉडल जटिलता सामान्यतः मॉडल अज्ञात फॅगसन्स के रूप में परिभाषित किया जाता है; इसलिए, दंडित अनुमान आम तौर पर अनुकूलन समस्या के हल के लिए है, इसका सामान्य रूप इस प्रकार है:

$$\left(\hat{\mu}, \hat{\beta}\right) = \left\{ \sum_i \left(Y_i - \mu - \sum_{j=1}^{p} X_{ij}\beta_j \right)^2 + \lambda J(\beta) \right\}$$

जहां λ एक नियमितीकरण पैरामीटर है जोकि मॉडल फिट की कमी तथा मॉडल जटिलता को नियंत्रित करता है। $\lambda = 0$ पर यह लीस्ट स्कवॉयर में बदल जाता है। कई आकलन की प्रक्रिया साहित्य में पहले से उपलव्ध है, और वे केवल पेनाल्टी चुनाव के लिए भिन्न होते हैं। रिज प्रतिगमन (आर आर) में (होरेल और केनार्ड 1970), पेनाल्टी सीधे प्रतिगमन गुणांक के वर्गों का योग के आनुपातिक होता है ,जैसे L2, $J(\beta) = \sum_{j=1}^{p} \beta_j^2$। सामान्यीकृत रूप में इसे ब्रिज प्रतिगमन के नाम से जाना जाता है(फ्रैंक और फ्राइडमैन 1993) तथा इसमें $J(\beta) = \sum_{j=1}^{p} \|\beta_j\|^{\gamma}$ को पेनाल्टी फँगसनके रूप में प्रयोग किया जाता है। रिज प्रतिगमन इसका एक विशेष रूप है जब $\gamma = 2$ होता है। $\gamma = 1$ होने पर ये पेनाल्टी फँगसनएक दूसरा रूप में बदल जाता है जिसे लिस्ट एबसेलुट एंगल एंड सलेक्शन ऑपरेटर (एल ए ए एस ओ) कहा जाता है। ये तरीके कई अनुप्रयोगों के लिए काफी उपयोगी होते हैं, लेकिन इन तरीकों की कई सीमाएं हैं।

बेसियन तरीके

मिऊविसीन आदि (2001) ने जिनोमिक सलेक्शन के लिए दो श्रेणीबद्ध मॉडल का प्रस्ताव रखा जोकि बेसियन दृष्टिकोण बेज ए तथा बेज बी पर आधारित था। किसी इन्डीविडुयल i के लिए मॉडल को इस प्रकार लिख सकते है:

$$Y_i = \mu + \sum_{j=1}^{p} X_{ij}\beta_j + e_i$$

जहाँ $i = 1(1)n,\ j = 1(1)p$ मार्कर स्थिति है , Y_i व्यक्ति विशेष i की फिनोटाइपिक वेल्यू है, μ एक$n \times 1$मिन वेक्टर है, X_{ij} इंसिडेंस मैट्रिक्स है

सामान्य में मॉडल को इस रूप में लिखा जा सकता है:

$$Y = \mu + \sum_{j=1}^{p} Z_j m_j + e$$

मॉडल मापदंडों के बारे में अनुमान पोस्टेरियर डिस्ट्रीवियुशनपर आधारित होता है। यह अच्छी तरह से ज्ञात है कि पोस्टेरियर डिस्ट्रीवियुशनको बेज प्रमेय का उपयोग करके प्रायर डिस्ट्रीवियुशनके साथ लाइकलीहूड फँग्सन के संयोजन से प्राप्त किया जा सकता है। बेज ए और बेज बी के बीच अंतर मुख्य रूप से उनके विचरण मापदंडों पर निर्भर होता है। बेज ए तरीका में प्रत्येक मार्कर पद के लिए एक ही विचरण का प्रयोग करते हैं। इसमें प्रायर डिस्ट्रीवियुशनके लिए हम लोग इंवर्टेड काइ-स्कवार डिस्ट्रीवियुशनका उपयोग करते हैं। बेज बी तरीका बेज ए कि तुलना में ज्यादा अच्छा माना जाता है।

बेसियन लासो

बेसियन लासो एक नई विधि है जिसे प्रतिगमन गुणांक का आकलन करने के लिये शुरू किया गया था (पार्क और कसेला, 2008) । बेसियन लासो में लासो विधि को बेसियन विश्लेषण के साथ कनेक्ट किया गया है। बेसियन लासो को लाइकलिहुड फंक्सन के साथ हाइरार्रिकल मॉडल का उपयोग कर जिनोमिक सलेक्शन में इस्तेमाल किया जाता है (डिलोस, कैम्पोस आदि 2009, 2010a, लांग आदि 2011)

$$f(Y|\mu, X, m, \sigma^2) \sim N(\mu + Xm, \sigma^2 I)$$

जहां Y $n \times 1$ डेटा वेक्टर है, μ समग्र मीन वेक्टर, m मार्कर प्रभाव का एक वेक्टर है, और X डिजाइन मैट्रिक्स जो m को Y से जोड़ता है, $N(\mu + Xm, \sigma^2 I)$ एक नोर्मल डिस्ट्रीवियुशनहै जिसका मध्य $\mu + Xm$ और विचरण $\sigma^2 I$ है। मार्कर $m_j's\ j = 1(1)p$ पर प्रायर डिस्ट्रीवियुशनके प्रभाव को लिखा जा सकता है $p(m_j|\tau_j^2) \sim N(0, \tau_j^2)$ तथा τ_j पर प्रायर डिस्ट्रीवियुशन $p(\tau_j|\lambda) \sim Exp(\lambda)$ है, जहां $Exp(\lambda)$ पैरामीटर λ के साथ घातीय डिस्ट्रीवियुशनको दर्शाता है।

निष्कर्ष

डीएनए मार्कर के व्यापक उपयोग सें प्रजनन कार्यक्रमों पर एक महत्वपूर्ण प्रभाव पड़ता है। पौधे और पशु प्रजनन में जेनेटिक इस्टीमेटेड ब्रीडींग वेल्यु की भविष्यवाणी एक केंद्रीय चुनौती है। सलेक्शन एक विशेषण समीकरण के आधार पर होना चाहिए। यह निश्चित रूप से सलेक्शन थ्रुपुट में सुधार करता है। यहाँ पर तुलनात्मक रूप से सबसे व्यापक रूप में इस्तेमाल किए जाने वाले जीनोमिक सलेक्शन के तरीकों का अध्ययन किया गया है। ये सभी तरीकें एडिटिव जेनेटिक आर्किटेक्चर के लिए अच्छे हैं। नान एडिटिव जेनेटिक आर्किटेक्चर के लिए साधारणतः नान-पैरामीट्रिक आधारित तरीको का उपयोग किया जाता है।

संदर्भ

- बर्नार्डो, आर. (2008). मॉलीकुलर मार्कर एंड सलेक्शन फॉर कॉम्पलेक्स ट्रैटस् इन प्लांटस्: लर्निंग फॉम लास्ट 20 इयर्स, 48: 1649-1664.
- बर्नार्डो, आर. (2012). ब्रीडिन्ग फॉर क्वानटेटिव ट्रैटस् इन प्लांटस्, स्टीरमा प्रेस ऊडबरी, एम एन.
- फ्रैंक, आई. ई. और फ्राइडमैन, जे. एच. (1993). ए स्टैटीस्टिकल विऊ ऑफ सम किमोमेट्रिकस् रिग्रेसन टूलस्. टेक्नोमेट्रिकस्, 35: 109-135.
- जिनोला, डी., फॅर्नांडो, आर. एल. और स्टेला, ए. (2006). जिनोमिक एसिसटेड प्रिडिक्सन ऑफ जेनेटिक वैल्यू विद सेमीपैरामेट्रिक प्रोसिजर्स, जेनेटिक्स, 173: 1761-1776.

- हस्टी, ट., तिव्शीरानी, आर. और फ्राइडमैन, जे (2009). द इलिमेंटस् ऑफ स्टैटिस्टिकल लर्निंग, 2 इडी. स्प्रिंगर।
- हेंन्डरसन, सी. आर. (1949) इस्टीमेट ऑफ चेंन्जेज इन हार्ड इनवरॉनमेंट. जे. डेयरी साइंस, 32: 706.
- हेंन्डरसन, सी. आर. (1953). इस्टीमेसन ऑफ भेरीएंस एंड कोभेरीएंस एंड कम्पोनेन्टस् बायोमेट्रिक्स, 9: 226-252.
- होरेल, ए. ई. और केनार्ड, आर. डब्ल्यू. (1970). रिज रिग्रेसन: बायस्ड इस्टिमेसन फॉर नॉन और्थोगोनल प्रोवलमस्, टेकनोमेट्रिक्स, 12: 55-67.
- लांग, एन, जियानोला, डी., रोजा, जी. जे. एम. और विगल, लालकृष्ण, ए. (2011). ऐप्लीकेसन ऑफ सपोर्ट भेक्टर रिग्रेसन टू जिनोम एसिसटेड प्रिडिक्सन ऑफ क्वानटेटिव ट्रैट्स थियोरेटिकल एंड एप्लाइड जेनेटिक्स, 123: 1065-1074.
- मियुविसिन, टी. एच. ई., हेस, बी. जे. और गोडार्ड, एम. ई. (2001). प्रिडिक्सन ऑफ टोटल जेनेटिक वेल्यू यूजिंग जिनोम वाइड डेन्स मर्कर मैप्स, जेनेटिक्स, 157: 1819-1829.
- टिवसिरानी, आर. (1996). रिग्रेसन श्रीन्केज एंड सलेक्शन वाया द लासो. जर्नल ऑफ रॉयल स्टैटीसटिकल सोसायटी, 58: 267-288.
- ज़ोय, एच. और हस्टी, टी. (2005). रेगुलेराइजेसन एंड वेरिएवल सलेक्शन वाया द इलासटिक नेट, जर्नल ऑफ द रॉयल स्ट सांख्यिकी रॉयल स्टैटीसटिकल सोसायटी, 67: 301-320.

डी.एन.ए. सिग्नेचर आधारित एस.एन.पी. और एस.टी.आर. मार्कर विश्लेषण

प्रस्तावना

आनुवंशिक संसाधनों के आण्विक लक्षण–वर्णन से संरक्षण हेतु निर्णय लेने में उद्देश्यपरकता तथा तार्किकता आती है। विभिन्न आण्विक मार्करों, विशेष रूप से माइक्रोसेटेलाइट, ए.एफ. एल.पी. व एस.एन.पी. द्वारा पादप, पशु, मत्स्य तथा सूक्ष्मजैविक आनुवंशिक संसाधनों का लक्षण–वर्णन किया जा रहा है जिसमें नाभिक जीनोम तथा माइटोकॉन्ड्रियाई जीनोम, दोनों शामिल हैं। इन आण्विक मार्करों में स्वनिर्मित 'आण्विक घड़ी' होती है जो विकासात्मक गति और पैमाने में गतिशील जननद्रव्य की विशिष्टता और भेदशीलता से संबंधित 'चित्र' या 'हस्ताक्षर' होते हैं। जैव सूचना विज्ञान ने न केवल जननद्रव्य के लक्षण–वर्णन में क्रांति लाई है बल्कि यह प्रजातियों की आण्विक पहचान की एक अपरिहार्य युक्ति भी सिद्ध हुआ है। जैव सूचना विज्ञान गैर–संवर्धित सूक्ष्मजीवों, पौधों, पशु व मत्स्य प्रजातियों की पहचान हेतु वर्गीकरण विज्ञान से लेकर सूक्ष्मजैविक मैटा–जीनोम विश्लेषण की सर्वाधिक सशक्ति युक्ति बन गया है। जीनोम विश्लेषण में हुई प्रगति से पशुओं, जीवाणुओं तथा विषाणु जीवों के बारे में अभूतपूर्व मात्रा में सूचना उपलब्ध हो रही है तथा इसमें रोगजनकों का पता लगाने व उनकी पहचान करने की भी बहुत क्षमता है। यहां एस.एन.पी. और एस.टी.आर. न्यूक्लियोटाइडों पर आधारित नाभिक अम्ल हस्ताक्षरों के विकास व उसके उपयोग की तार्किक युक्ति का वर्णन किया गया है। इस प्रकार के आण्विक मार्करों के वर्गीकरण तथा पूर्वानुमान की अन्य जीवसूचना विज्ञानी युक्तियों की भी चर्चा की गई है।

प्रजातियों की डी.एन.ए. बारकोडिंग तथा इसका उद्भव

डी.एन.ए. बारकोडिंग एक वर्गीकरणविज्ञानी विधि है जिसमें किसी विशेष प्रजाति की पहचान के लिए जीव के माइटोकॉन्ड्रियाई डी.एन.ए. में छोटे आनुवंशिक मार्कर का उपयोग किया जाता है। यह अपेक्षाकृत एक सरल संकल्पना पर आधारित है : अधिकांश यूकैरियोट कोशिकाओं में माइटोकॉन्ड्रियाई और माइटोकॉन्ड्रियल डी.एन.ए. (एम.टी.डी.एन.ए.) होता है तथा इसकी अपेक्षाकृत तीव्र उत्परिवर्तन दर होती है जिसके कारण प्रजातियों के बीच एम. टी.डी.एन.ए. क्रमों में उल्लेखनीय विविधता होती है और सैद्धांतिक रूप से प्रजातियों में अपेक्षाकृत कम भिन्नता होती है। आरंभ में एक 648–बीपी क्षेत्र के साइटोक्रोम सी–ऑक्सीडेस उप इकाई ५ जीन (सी.ओ.आई.) को एक सक्षम 'बारकोड' के रूप में प्रस्तावित किया गया था।

विकासात्मक संबंधों की खोज करने के लिए न्यूक्लियोटाइड क्रम में विविधता का उपयोग करना एक नई संकल्पना है। कार्ल वोइजे ने आर्की की खोज के लिए राइबोसोमी आर.एन. ए. (rRNA) में क्रम में मौजूद भेदों का उपयोग किया था जिससे आगे चलकर विकासात्मक वृक्ष तथा आण्विक मार्करों (जैसे एलोजाइम, rDNA, तथा mtDNA vage) की प्राप्ति हुई। डी. एन.ए. बारकोड प्रजाति की पहचान के लिए 'बारकोड' उपलब्ध कराने हेतु जीनोम के किसी विशेष क्षेत्र से छोटे डी.एन.ए. क्रम के उपयोग के माध्यम से सम्पन्न की जाने वाली क्रियाविधि उपलब्ध कराता है। वर्ष 2003 में पॉल डी.एन. हर्बर्ट, ग्यूएल्फ विश्वविद्यालय, ओंटारियो, कनाडा ने डी.एन.ए. बारकोडिंग की एक ऐसी पब्लिक लाइब्रेरी का संकलन प्रस्तावित किया था जिसे नामित नमूनों से जोड़ा जा सकता है। यह लाइब्रेरी 'प्रजाति की पहचान के लिए एक नई

मुख्य कुंजी उपलब्ध कराएगी जिसकी शक्ति वर्गीकरण की कवरेज के साथ बढ़ेगी तथा इससे तीव्र लेकिन सस्ता अनुक्रमण या सीक्वेंसिंग हो सकेगा।'

प्रजाति बारकोड द्वारा पक्षियों की पहचान

वर्गीकरण द्वारा स्थापित और डी.एन.ए. बारकोडिंग द्वारा व्याख्या की गई परंपरागत प्रजाति सीमाओं के बीच के संबंध का पता लगाने के प्रयास में हैबर्ट और उसके साथियों ने उत्तर अमेरिका में 667 पक्षी प्रजातियों के डी.एन.ए. बारकोड क्रमबद्ध किए (हैबर्ट और साथी, 2004)। यह पाया गया कि इन 260 प्रजातियों के विभिन्न सी.ओ.आई. क्रम थे। 130 प्रजातियों का दो या इससे अधिक नमूनों द्वारा प्रतिनिधित्व हुआ। इन सभी प्रजातियों में सीओआई क्रम या तो समान थे या उसी प्रजाति के क्रमों के सर्वाधिक समान थे। प्रजातियों के बीच सीओआई विविधताएं 7.93 प्रतिशत थीं, जहां प्रजातियों के बीच विविधता औसतन 0.43 प्रतिशत थी। चार मामलों में गहरी अंतर–प्रजातीय विविधता पाई गई जिससे संभावित नई प्रजाति का संकेत मिला। इन चार पॉलीटाइपिक प्रजातियों में से तीन का वर्गीकरण कुछ वैज्ञानिकों ने पहले से ही कर दिया है। हैबर्ट और उनके साथियों, (2004) के इस कार्य से इन विचारों को बल मिला तथा डी.एन.ए. बारकोडिंग के लिए मामला मजबूत हुआ। उन्होंने नई प्रजाति को परिभाषित करने के लिए मानक क्रम थ्रेशहोल्ड भी प्रस्तावित किया। यह थ्रेशहोल्ड जो तथाकथित 'बारकोडिंग गैप' है। अध्ययन के अंतर्गत समूह हेतु माध्य अंतरा प्रजाति विभिन्नता 10 गुनी होने के रूप में परिभाषित किया गया है।

डी.एन.ए. बार कोड द्वारा क्रिप्टिक प्रजातियों को असीमित करना

डी.एन.ए. बार कोडिंग की कुशलता से संबंधित अन्य प्रमुख अध्ययन में उत्तर–पश्चिम कोस्टा अधिकार में डी–गुआना कास्ट के क्षेत्र संरक्षण (ए.सी.जी.) पर नियो ट्रॉपिकल स्कीपर बटरफ्लाई, *एस्ट्राप्टेस्फल्गरेटर* पर ध्यान केन्द्रित किया गया। यह प्रजाति पहले से ही क्रिप्टिक प्रजाति जटिलता के लिए विख्यात है क्योंकि इसमें कठोर आकृतिविज्ञानी भेद होते हैं तथा यह खाद्य पौधों की इल्ली की असामान्य रूप से बड़ी किस्म है। तथापि, प्रजाति को पूर्णतः असीमित करने के लिए वर्गीकरण वैज्ञानिकों को अनेक वर्ष लगेंगे। हैबर्ट और साथियों ने (2004बी) एसीजी से 484 नमूनों के सीओआई जीन को क्रमबद्ध किया। इस नमूने में 'प्रत्येक खाद्य पौधे की प्रजाति पर पाले गए कम से कम 20 वैयक्तिक, वयस्कों के अंतिम अवस्था वाले व मध्यवर्ती व इल्लियों में रंग विविधता वाले प्रतिनिधियों को शामिल किया गया' जो उन तीन पारिस्थितिक प्रणालियों से थे जहां *एस्ट्राप्टेस्फल्गरेटर* पाया गया था। हैबर्ट तथा साथियों (2004b) ने निष्कर्ष निकाला कि *एस्ट्राप्टेस्फल्गरेटर* की उत्तर–पश्चिम कोस्टारिका में दस विभिन्न प्रजातियां हैं। इससे यह उजागर होता है कि डी.एन.ए. बार कोडिंग विश्लेषणों के परिणाम अन्वेषणकर्ताओं द्वारा प्रयुक्त विश्लेषणात्मक विधियों के चयन पर निर्भर करते हैं। इसलिए डी.एन.ए. बार कोडों का उपयोग करके क्रिप्टिक प्रजातियों को असीमित करने की प्रक्रिया किसी अन्य वर्गीकरण विज्ञान की तुलना में विषयपूरक हो सकती है।

प्रजाति डी.एन.ए. बार कोड द्वारा पुष्प पौधों की पहचान

क्रैस और साथियों (2005) ने सुझाया था कि सी.ओ.आई. क्रम का उपयोग पौधों की अधिकांश प्रजातियों के लिए उचित नहीं है क्योंकि पशुओं की तुलना में उच्च पादपों में सी ऑक्सीडेस

जीन विकास की दर अत्यधिक धीमी होती है। पुष्प पौधों की डी.एन.ए. बार कोडिंग में उपयोग के लिए जीनोम के सर्वाधिक उपयुक्त क्षेत्र का पता लगाने के लिए अनेक प्रयोग किए गए।

उचित आनुवंशिक स्थल के लिए तीन आधार निर्धारित किए गए:

1. उल्लेखनीय प्रजाति स्तर की आनुवंशिक भिन्नता और विविधता
2. उचित छोटी क्रम लंबाई ताकि डी.एन.ए. निष्कर्षण और आवर्धन में सुविधा हो, तथा
3. वैश्विक प्राइमर विकसित करने के लिए संरक्षित फ्लैंकिंग स्थलों की उपस्थिति।

इन प्रयोगों के पूरा होने पर परिणाम के रूप में क्रैस और साथियों (2005) ने नाभिक आंतरिक ट्रांसक्राइब किया गया अंतराल प्रस्तावित किया तथा पुष्प पौधों के सक्षम डी.एन.ए. बार कोड के रूप में प्लास्टिड trnH-psbAintergenic spacer को भी प्रस्तावित किया। परिणामों से यह सुझाव मिलता है कि डी.एन.ए. बार कोडिंग 'मुख्य कुंजी' होने की बजाय 'मुख्य कुंजी का छल्ला' हो सकता है जिसमें जीवन की विभिन्न जाति आती हैं जिन्हें विभिन्न कुंजियों की आवश्यकता होती है।

कवकों के प्रभेदों की पहचान

पक्सीनिया ग्रेमेनिस जो तना रतुआ का एक कारक एजेंट है, से छोटे अनाजों (गेहूं, जौ, जई और राई) में विश्वभर से गंभीर रोग होता है। *पी.ग्रैमिनिस* रतुआ कवक (यूरेडिनेल) का प्रथम क्रमबद्ध किया गया प्रतिनिधि है जो एक अनिवार्य पादप रोगजनक है। रतुआ कवक की 7000 से अधिक प्रजातियां हैं और यह पादप रोगजनकों का सर्वाधिक विनाशकारी समूह है। गेहूं का तना रतुआ उन सभी क्षेत्रों में गंभीर समस्या बन गया है जहां गेहूं की खेती की जाती है तथा इससे उत्तर अमेरिका में प्रमुख महामारी फैली थी। वर्ष 1999 में *पी. ग्रेमिनिस* का एक नया अति उग्र प्रभेद टी.टी.के.एस. (यूजी 99) युगांडा में पहचाना गया और इसके बाद फैलता गया जिससे केन्या और इथोपिया में महामारी व्याप्त हुई।

जैव सूचना विज्ञान प्रजाति तथा प्रभेदों की पहचान में और इसके साथ ही पूरे विश्व में इसकी गतिकी को समझने में महत्वपूर्ण भूमिका निभा सकता है। परपोषी और परजीवी दोनों पर वृहत आंकड़े सृजित किए गए हैं जो नवीनतम आण्विक या जैवप्रौद्योगिकीय युक्तियों से तैयार हुए हैं और इनका जैव सूचना विज्ञानी युक्तियों द्वारा आसानी से विश्लेषण किया जा सकता है। हमारी वार्ता *पी. ग्रेमिनिस* की यूजी 99 पर केन्द्रित रहेगी। इसके जीनोम का किस प्रकार कवकीय प्रजातियों की गति का पता लगाने के लिए उपयोग किया जा सकता है तथा जैव सूचना विज्ञानी युक्तियां किस प्रकार यूजी 99 की पहचान सहित *पी. ग्रेमिनिस* के प्रभेद की पहचान में सहायक सिद्ध हो सकती हैं।

पालतू प्रजातियों तथा पशु नस्लों के डी.एन.ए. आधारित सिग्नेचर

माइटोकॉन्ड्रियाई डी.एन.ए. मार्कर घरेलू पशुओं में सफल सिद्ध हुआ है और इसका उपयोग विशिष्ट रूप से मांस की पहचान, वन्य पशुओं के शिकार, डेरी दूध, डेरी उत्पादों (जैसे चीज) में मिलावट का पता लगाने के लिए किया जा सकता है। नस्लों के लिए प्रयुक्त होने वाले वर्तमान मार्कर लगभग एस.टी.आर. हैं लेकिन अभी हाल ही में एस.एन.पी. आधारित चिप नस्ल के सिग्नेचर का निर्धारण करने और इसके साथ ही जनकता तथा वंशावली में अधिमिश्रण का पता लगाने के मामले में अति शक्तिशाली युक्ति सिद्ध हुई है।

नस्लों के एस.टी.आर. आधारित सिग्नेचर

नस्लों के आण्विक लक्षण–वर्णन के संबंध में विभिन्न वैज्ञानिकों मंचों पर सामान्यतः यह प्रश्न पूछा जाता है कि क्या रक्त, वीर्य, बाल, रक्त के धब्बे, शव आदि के नमूने से पशुधन नस्ल की पहचान की जा सकती है। पिछले कुछ वर्षों के दौरान आण्विक आनुवंशिक विदों द्वारा विश्व में इस प्रश्न का उत्तर देने के अनेक प्रयास किए गए हैं। कुछ अध्ययन स्पेन, पुर्तगाल, फ्रांस में गोपशुओं की विभिन्न नस्लों की नस्ल पहचान तथा नस्ल–विशिष्ट आनुवंशिक/डी. एन.ए. सिग्नेचर हेतु प्रौद्योगिकी विकसित करने में सफल हुए हैं; जैसे नॉर्वे में घोड़ों, स्पेन में भेड़, केन्या में ऊंट, आदि जैसे पशुओं के मामले में कार्य हुआ है। इन अध्ययनों में नस्ल के प्रमाणीकरण की सटीकता का अंश अत्यधिक उच्च था जो 95 प्रतिशत से 99 प्रतिशत के बीच रहा।

ये विधियां नामतः (प) आवर्तता विधि (पीटकाउ और साथी, 1995), (पप) बायेसियन विधि (रान्नाला और साथी, 1997) और (पपप) दूरी विधि (गोल्ड स्टेइन और साथी, 1995) नस्ल विशिष्ट सिग्नेचर विकसित करने के लिए उपयोग में लाई गई हैं। इनमें से बायेसियन विधि माइक्रोसेटेलाइट आंकड़ों की झ99 प्रतिशत विश्वास सीमा के साथ सर्वाधिक सटीक रिपोर्ट की गई है (कारेंडर और साथी, 2003; (बुस्तामांग्ते और साथी, 2003)।

विदेशों में हाल के वर्षों में पशुओं की कुछ प्रजातियों के आनुवंशिक सिग्नेचर विकसित करने के कुछ प्रयास हुए हैं। जहां किसी पशु के अपमिश्रित नस्ल के होने के कारण उसे या उस विशेष नस्ल को कोई वैयक्तिक पहचान प्रदान करना कठिन हो जाता है। ऐसी संदेहास्पद नस्ल की पहचान के मामलों में नस्ल संकर सूचकांक संबंधी अध्ययन किए गए हैं। अतः साहित्य की समीक्षा दो शीर्षकों के अंतर्गत की गई है (1) नस्ल विशिष्ट सिग्नेचरों/प्रोफाइलों का विकास और (2) नस्ल संकर सूचकांक का विकास।

नस्लों के डी.एन.ए. सिग्नेचर पर आधारित एस.एन.पी. चिप

जापान में जापानी ब्लैक तथा होल्सटेइन गोपशुओं को मांस के लोकप्रिय स्रोत के रूप में सराहा जाता है तथा आस्ट्रेलिया व संयुक्त राज्य अमेरिका से आयात किए गए गोमांस की भी मांस उद्योग में काफी मांग है। बी.एस.ई. के प्रकोप के बाद मिथ्या बिक्री की समस्या उत्पन्न हुई है : आयातित गोमांस को अक्सर उपभोक्ताओं की चिंता के कारण घरेलू गोमांस के रूप में गलत ढंग से लेबल किया गया है। अतः खाद्य सुरक्षा के लिए जापानी तथा आयातित गोपशुओं के बीच सही भेद करने की एक विधि की आवश्यकता है। एस.एन.पी. 50के आधारित चिप जापानी तथा अमेरिकी गोपशु के बीच मार्करों में भेद कर सकती है। ऐसी रिपोर्ट है जहां संयुक्त राज्य अमेरिका विशिष्ट मार्कर (बीआईएसएनपी 7, बीआईएसएनपी 15, बीआईएसएनपी 21)

बीआईएसएनपी 23 और बीआईएसएनपी 26) विकसित किए गए हैं जिनकी युग्मविकल्पता आवर्तताएं 0.102 (बीआईएसएनपी 15) से 0.250 (बीआईएसएनपी 7) के बीच है जिनका औसत 0.184 है। इन पांच मार्करों के सम्मिलित उपयोग से 0.858 की पहचान संभाव्यता के साथ जापानी और अमेरिकी गोपशुओं में भेद करना संभव होगा। परिणामों से नस्ल पहचान हेतु मार्करों के विकास के लिए गोपशु 50 के एस.एन.पी. ऐरे की एक सशक्त युक्ति के रूप में क्षमता का संकेत मिलता है। ये मार्कर जापान में मिथ्या गोमांस संबंधी डिस्प्ले से बचाव करने में अपना योगदान देंगे (स्यूकाव और साथी, 2010, सासाजाकी और साथी, 2011)।

पादप किस्म का डी.एन.ए. आधारित सिग्नेचर, उदाहरण– बासमती चावल

बासमती चावल की एक विशिष्ट पांडन–जैसी (*पोंडानुसएमेरीलीफोलियस* की पत्ती) गंध होती है जो यौगिक 2–एसिटाइल–1–पाइरोलीन की गंध से उत्पन्न होती है। नकली तथा असली परंपरागत बासमती चावल में भेद करने में कठिनाई आती है तथा धोखेबाज व्यापारियों द्वारा बासमती चावल का मूल्य उच्च होने के कारण, असली चावल में नकली चावल की मिलावट की जाती है। उपभोक्ताओं तथा व्यापार के हितों की रक्षा के लिए मनुष्यों में डी.एन. ए. फिंगर प्रिंटिंग के समान एक पीसीआर आधारित मूल्यांकन विधि विकसित की गई है जिससे मिलावटी तथा असली बासमती प्रभेदों की पहचान की जा सकती है। मिलावट के लिए इसकी पहचान सीमा 1 प्रतिशत या इससे अधिक है तथा त्रुटि दर $\pm$ 1.5 प्रतिशत है। बासमती चावल के निर्यातक बासमती चावल की अपनी खेपों के लिए डी.एन.ए. परीक्षणों पर आधारित 'शुद्धता प्रमाण–पत्रों' का उपयोग करते हैं। इसका विकास डी.एन.ए. फिंगर प्रिंटिंग एवं डाइग्नोस्टिक्स के लिए लैबिंडिया नामक केन्द्र ने किया है जो एक भारतीय कंपनी है तथा इसने बासमती चावल में मिलावट का पता लगाने के लिए किटें जारी की हैं। यह बासमती चावल में मिलावट के लिए विश्व की प्रथम एकल–ट्यूब वाली, मल्टी प्लैक्स (8 माइक्रोसेटेलाइट स्थलों को आवर्धित करने के लिए) युक्त माइक्रोसेटेलाइट मूल्यांकन–आधारित किट है।

बासमती वैरिफाइलर^{TM} किट आण्विक मूल्यांकन के माध्यम से बासमती चावल की प्रमाणिकता को स्थापित करने हेतु विकसित किया गया विश्व का प्रथम उत्पाद है। इस किट में पीसीआर आवर्धन तकनीक का उपयोग किया जाता है जो साधारण क्रम दोहराव (एस.एस.आर.) पर आधारित है और इससे बासमती जीन प्ररूपण के लिए एकल सर्वाधिक विभेदनशील मूल्यांकन करना संभव है।

मछलियों के डी.एन.ए. आधारित बार–कोडित हस्ताक्षर

वार्ड और साथियों (2015) ने अपने अनुसंधान पत्र में मछली प्रजातियों की पहचान में कॉक्स 1 सीक्वेंसिंग या 'बार कोडिंग' की क्षमता का वर्णन किया है। इस अध्ययन में 207 मत्स्य प्रजातियों जिनमें से अधिकांश आस्ट्रेलियाई समुद्री मछलियां थीं, को माइटोकॉन्ड्रियाई साइटोक्रोम ऑक्सीडेस उप इकाई ५ जीन (कॉक्स 1) के 655 बीपी क्षेत्र के लिए अनुक्रमित (बार कोडित) किया गया। अधिकांश प्रजातियों का प्रतिनिधित्व अनेक नमूनों के द्वारा हुआ तथा कुल 754 क्रम सृजित किए गए। टेलीओस्टों की 143 प्रजातियों का जीसी अंश शार्कों तथा रे मछलियों की 61 प्रजातियों की तुलना में उच्च था (47.1 प्रतिशत बनाम 42.2 प्रतिशत)। इसका कारण टेलिओस्टों में कोडोन स्थल 3 पर उच्च वीसी अंश का होना है (41.1 प्रतिशत बनाम 29.9 प्रतिशत)। रे मछलियों में शार्कों की तुलना में उच्चतर जीसी था (44.7 प्रतिशत बनाम 41.0 प्रतिशत)। इसका कारण भी पहले वाली मछली की तीसरी कोडोन स्थिति में जीसी का उच्चतर होना था। प्रजातियों, वंश, कुल , गण तथा वर्ग किमुरा में दो प्राचल (के2पी) की औसत दूरियां क्रमशः 0.39 प्रतिशत, 9.93 प्रतिशत, 15.46 प्रतिशत, 22.18 प्रतिशत और 23.27 प्रतिशत थीं। सभी प्रजातियों में उनके कॉक्स 1 क्रम के द्वारा भेद करना संभव हुआ। यद्यपि प्रत्येक दो प्रजाति के एकल वैयक्तिक हैप्लोटाइप गुण वाले थे। यद्यपि डी.एन. ए. बारकोडिंग का उद्देश्य प्रजातियों की पहचान प्रणालियों का विकास करना है तथापि इन आंकड़ों में कुछ फाइलोजेनेटिक संकेत मिलना बहुत स्वभाविक है। सभी 754 क्रमों के लिए पड़ोस से जुड़ने वाले वृक्ष में चार मुख्य गुच्छे दिखाई दिए : काइमीरिड्स, रे, शार्क और

टेलियोस्ट वंशों में प्रजातियों को अनिवार्य रूप से समूहित किया गया और ऐसा ही कुलों में वंशों के साथ भी किया गया। तीन वर्गीकरण विज्ञानी समूह – वंश स्क्वालस की डॉगफिश, कुल प्लेटीसिफेलिडी की फ्लैटहैट्स तथा वंश थुन्नुस की तूना को और निकटता से जांचा गया। बूटस्ट्रैपिंग के पश्चात् स्पष्ट हुए क्लैड अपेक्षाओं के अनुकूल थे। परिचालनात्मक वर्गीकरण इकाइयों से वैयक्तिकों को वैयक्तिक क्लैड द्वारा निर्मित थ के माध्यम से स्क्वालस प्रजाति ठ के रूप में नाम दिया गया। इनमें से प्रत्येक के लिए समर्थनकारी आकृतिविज्ञानी प्रमाण भी जुटाए गए हैं। इस शोध पत्र को डी.एन.ए. आधारित मत्स्य सिग्नेचर के लिए व्यापक रूप से उद्धृत किया जाता है।

आण्विक आंकड़ों के वर्गीकरण व पूर्वानुमान के लिए विभिन्न जैव सूचना विज्ञानी युक्तियां

जीनोम विश्लेषण प्रौद्योगिकी में हुई प्रगतियां से पशुओं, जीवाणुओं तथा विषाण्विक जीवों के बारे में अभूतपूर्व सूचना उपलब्ध हो रही है तथा इसमें रोगजनकों का पता लगाने व उनकी पहचान करने की बहुत क्षमता है। लेख के इस भाग में एस.एन.पी. और एस.टी.आर. न्यूक्लियोटाइडों पर आधारित नाभिक अम्ल सिग्नेचरों के विकास तथा अनुप्रयोग की तर्कसंगत युक्ति का वर्णन किया गया है।

एस.एन.पी. (जैसे, अनुक्रमण तथा सार्वजनिक डेटाबेसों) का ध्यान न भी रखा जाए तो भी लक्षित जीव तथा उसके निकटतम पड़ोसियों से जब एक बार एस.एन.पी. एकत्र कर लिए जाते हैं तो ऐसे एस.एन.पी. की पहचान करना आवश्यक हो जाता है जो किसी प्रजाति या प्रभेद की पहचान में उपयोगी सिद्ध हो सकते हैं। इस युक्ति में एस.एन.पी. मार्करों के डेटाबेस का उपयोग किया गया है, ताकि विकासात्मक क्लेडों की पहचान के लिए फाइलोजेनेटिक विश्लेषण किया जा सके तथा एस.एन.पी. को परिभाषित किया जा सके। विशाल आंकड़ों के भंडारण की क्षमता जिससे आंकड़ों तक पहुंचने में सुविधा होगी, से स्वचालित एस.एन.पी. पूर्वानुमान (मानवीय हस्तक्षेप में कमी लाकर) सिग्नेचर का रेखांकन करना संभव हुआ है और इससे जटिल प्रश्नों का उत्तर देने में सुविधा हुई है। स्थानीय संसाधनों के रूप में अनेक डेटाबेस विद्यमान हैं तथापि, यूकैरियोटिक एस.एन.पी. डेटा से युक्त कुछ वैश्विक डेटाबेस भी स्थापित किए गए हैं (जैसे डी.बी.एस.एन.पी.)। इस प्रकार के वैश्विक डेटाबेस सूक्ष्मजैविक एस.एन.पी. डेटा के लिए विकसित नहीं हुए हैं। एस.एन.पी. खोज तथा फाइलो जेनेटिक विश्लेषण के लिए सृजित प्रत्येक डेटाबेस की विषय–वस्तु व संरचना भिन्न होगी जिसका निर्धारण आंकड़ों के उपयोगों द्वारा करना संभव होगा। डेटाबेस डिजाइन करने का कोई एक मात्र सही तरीका नहीं है। तथापि, तथ्य यह है कि संचार कर्मी के लिए विभिन्न बहुरूपण डेटाबेसों को ही स्वीकार करना होगा बल्कि आंकड़ों के विश्लेषण के लिए अनिवार्य सूचना भी प्रदान करनी होगी। चार मुख्य अनिवार्य तत्व परिभाषित किए गए हैं और इनमें शामिल हैं:

- एक अनूठा एस.एन.पी. पहचानकर्ता (युग्मविकल्पी)
- आंकड़ा स्रोत (जैसे प्रयोगात्मक या कम्प्यूटेशनल)
- युग्मविकल्पी और युग्मविकल्पियों को फ्लैंक करने वाला क्रम

यूकैरियोटी एस.एन.पी. आंकड़ों के भंडारण व विश्लेषण के लिए अनेक डेटाबेस सृजित किए गए हैं, इनमें से कुछ वृहत हैं और कुछ जीनोमवार हैं तथा अन्य विशेषज्ञतापूर्ण या स्थल–विशिष्ट हैं। दोनों प्रकार के डेटाबेस अनिवार्य हैं। वृहत डेटाबेस से बहुरूपण की दृष्टि से जीनोमवार आंकड़े उपलब्ध होंगे जो प्रभेद निर्धारण व पहचान के लिए आदर्श होंगे।

स्थल–विशिष्ट डेटाबेस से किसी विशेष स्थल पर बहुरूपों का अधिक गहरा व अधिक स्पष्ट दृश्य प्राप्त होगा। किसी डेटाबेस में ऐसी सटीक सूचना शामिल होनी चाहिए जिसका उपयोग डाउन स्ट्रीम विश्लेषणों के लिए किया जा सके और जिसमें अन्य डेटाबेसों के साथ एकीकृत होने की क्षमता हो। एस.एन.पी. से संबंधित कुछ अतिरिक्त सूचना भी डेटाबेस में कार्यान्वित की जानी चाहिए। डेटाबेस तथा इससे सम्बद्ध पाइपलाइन को अनेक स्रोतों से आंकड़ों को संसाधित व भंडारित करने में सक्षम होना चाहिए, न कि उसे केवल बाह्य क्रम डेटाबेसों के लिए एक अनुक्रमण के यंत्र के रूप में इस्तेमाल किया जाना चाहिए (जैसे जीन बैंक, डीबीईएसटी)। डेटाबेस के द्वारा किसी जीव को ट्रैक करना संभव होना चाहिए तथा उससे यह स्पष्ट होना चाहिए कि कौन सा एस.एन.पी. जीनोम–, जीन– और एक्सॉन–विशिष्ट सूचना जो एस.एन.पी. से जुड़ी हो, के लिए सार्थक है। डाउन स्ट्रीम विश्लेषण के लिए क्रमों की फ्लैंकिंग ही पर्याप्त नहीं है बल्कि इसके लिए संदर्भ क्रम भी होना चाहिए। गुणवत्ता सुनिश्चित करने के उद्देश्य तथा सामान्य आंकडों के विश्लेषण हेतु उपयोगी सूचना में ऐसा एल्गोरिथ्म शामिल है जिसके द्वारा एस.एन.पी. की खोज हुई थी और यह भी ज्ञात किया जाना चाहिए कि क्या इसका प्रयोग द्वारा सत्यापन किया गया था या सत्यापन न करके इसका कम्प्यूटेशन के माध्यम से पूर्वानुमान लगाया गया था। साथ ही उस विधि का भी पता लगाया जाना चाहिए जिसके द्वारा इसका सत्यापन हुआ था (जैसे जीनप्ररूपण मूल्यांकन या अनुक्रमण)। एस.एन.पी. के प्रकार को औसत युग्मविकल्पी आवर्तता के साथ शामिल किया जाना चाहिए (जैसे समयुग्मज या विषमयुग्मज)। उपयोगी सूचना जैसे क्रम के संदर्भ में एस. एन.पी. से संबंधित स्थिति, कोंटिंग या एम्प्लीकॉन और एस.एन.पी. मूक है या रोगजनक, को भी शामिल किया जाना चाहिए। सिग्नेचर विकास की आवश्यकताओं को पूरा करने के लिए एक तर्कसंगत डेटाबेस सृजित किया गया है, ताकि एस.एन.पी. को इससे संबंधित सूचना भंडारित की जा सके और डाउन स्ट्रीम मूल्यांकन की विधि विकसित हो सके। एस.एन.पी. खोज तथा मूल्यांकन डिजाइन से विशिष्ट सूचना को डेटाबेस टेबलों या इंटाइटिस में तर्कसंगत रूप से भंडारित किया जाना है ताकि एस.एन.पी. और संबंधित आंकड़ों संबंधी जटिल शंकाओं का समाधान किया जा सके। विशेष रूप से एस.एन.पी. टेबल में एस.एन.पी. स्थल युग्मविकल्पियों को शामिल करना तथा मूल्यांकन डिजाइन के लिए 5′ और 3′ फ्लेंकिंग क्रम लिए जाने चाहिए। जीन, एक्सॉन तथा प्रोजेक्ट से संबंधित सूचना को डाउन स्ट्रीम विश्लेषण जैसे समष्टि संबंधी अध्ययनों में सुविधा प्रदान करने के लिए भंडारित किया जाता है। एल्गोरिथ्म विशिष्ट श्रेणी मानों और विधियों को भी शामिल किया जाता है जिनसे अन्वेषणकर्ता प्रत्येक एस.एन.पी. की वास्तविक गुणवत्ता का मूल्यांकन करता है। एस.एन.पी. टेबल डेटाबेस की केन्द्रीय मूल आवश्यकता है। प्रत्येक एस.एन.पी. से वह नाम जुड़ा होता है जहां प्रत्येक एस.एन.पी. को एक से अधिक नाम भी दिया जा सकता है। प्रत्येक एस.एन.पी. को एक या इससे अधिक संदर्भ क्रमों से भी सम्बद्ध किया जा सकता है। संदर्भ क्रमों के अनेक उद्देश्य हैं जिनमें शामिल हैं :

- ✓ पी.सी.आर. प्राइमर डिजाइन के लिए टैम्पलेट के रूप में कार्य करना
- ✓ किसी एस.एन.पी. के चारों ओर फ्लैंकिंग क्रम उपलब्ध कराना
- ✓ सटीक एसेम्बली सुनिश्चित करने के लिए फार्प एसेम्बली में शामिल किया जाना

संदर्भ क्रम कार्यात्मक स्वचालन का आरंभ बिंदु भी उपलब्ध कराते हैं। संदर्भ क्रम किसी नाम, जीन बैंक प्रविष्टि या जीआई संख्या,विवरण व क्रम से युक्त होता है। एस.एन.पी. पूर्वानुमान

के लिए प्रयुक्त एम्प्लीकॉनों का अनुक्रमण किया जाता है। एम्प्लीकॉन से संबंधित सूचना में शामिल है प्रत्येक एम्प्लीकॉन का नाम और विवरण, इसके आवर्धन के लिए प्रयुक्त किए गए प्राइमर और इसका अपेक्षित आकार। यद्यपि यह डेटाबेस उच्च यूकैरियाटों तथा उनके विषाणुओं के लिए डिजाइन किया गया है तथापि यह आंकड़ा संबंध प्रोकैरियोटिक एस.एन. पी. आंकड़ों के लिए भी वही रहेगा। एस.एन.पी. मार्कर डेटाबेस डाउन स्ट्रीम सिग्नेचर विकास तथा मूल्यांकन डिजाइन संबंधी क्रियाओं के लिए वांछित सूचना के भंडारागार के रूप में भी कार्य करता है। प्रत्येक एस.एन.पी. को खोजने, क्रम आंकड़ा विश्लेषण, एस.टी.आर. आंकड़ा विश्लेषण के लिए महत्वपूर्ण वे सूचना विज्ञानी युक्तियां जो डी.एन.ए. सिग्नेचरों पर आधारित एस.एन.पी./एस.टी.आर. विकसित करने हेतु उपयोग में लाई जाती हैं, उनसे संबंधित प्रोटोकॉल व मूल सूचना नीचे दर्शाई गई हैं :

Gene class 2.0

गोपशुओं तथा भैंसों की विभिन्न नस्लों को दिए जाने हेतु एकल न्यूक्लियोटाइड पॉलीमॉर्फिज्म (एस.एन.पी.) की प्रभावशीलता पर अनेक एस.एन.पी. के विश्लेषणों द्वारा अन्वेषण किए गए हैं। नस्ल का निर्धारण बायेसियन तथा आवर्तता विधियों की तुलना के द्वारा किया जाता है जो स्ट्रक्चर 2.2 तथा जीन क्लास 2 सॉफ्टवेयर कार्यक्रमों में लागू की गई हैं। ज्ञात वैयक्तिकों के पुनरावंटन के लिए एस.एन.पी. का उपयोग उनकी नस्लों के उद्गम हेतु किया जाता है तथा अज्ञात वैयक्तिकों हेतु एस.एन.पी. के निर्धारण संबंधी कार्य का परीक्षण किया जा चुका है। इसका उदाहरण भैंस में जीन वर्ग 2 के रूप में दिया जाता है जिसमें भैंस की नस्लों के संदर्भ तथा अज्ञात आंकड़े उपलब्ध हैं (चित्र 1 और 2)। इससे संबंधित चरण निम्नानुसार हैं :

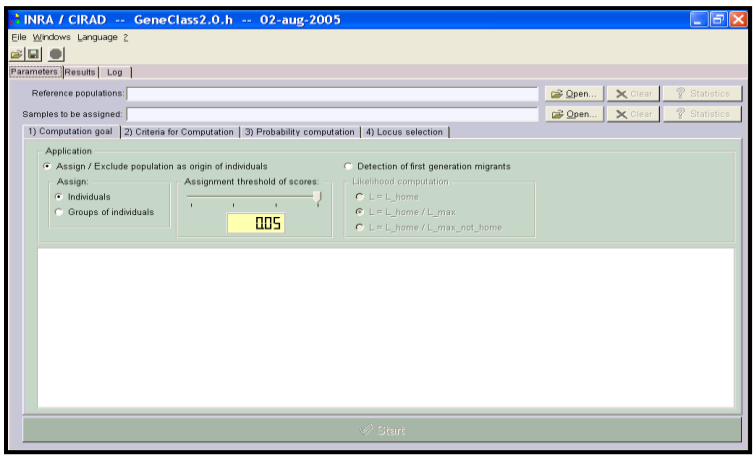
चरण 1: GeneClass2.0 सॉफ्टवेयर डाउनलोड करें (http://www.montpelier.inra.fr/URLB/geneclass/genclass.html पर निःशुल्क उपलब्ध)

चरण 2: संदर्भ तथा अज्ञात नमूनों के लिए डेटा फाइलें तैयार करना

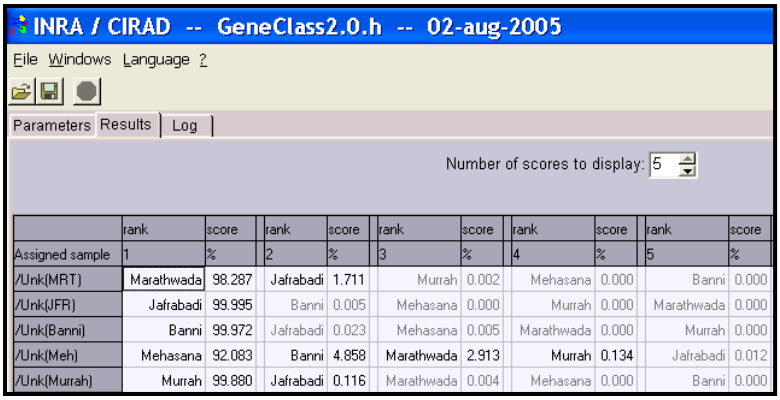
चरण 3: सॉफ्टवेयर का मेन विंडो खोलना (चित्र 1) तथा दोनों फाइलों को इम्पोर्ट करना।

चरण 4: कम्प्यूटेशनल लक्ष्य, कम्प्यूटेशन के लिए आधार, संभाव्यता कम्प्यूटेशन तथा चयन आधार जैसे प्राचलों का चयन करना।

चरण 5 : स्टार्ट बटन को क्लिक करके हम परिणाम देख सकते हैं (चित्र 2) तथा अंततः परिणामों की व्याख्या की जा सकती है।



चित्र 1: Gene class 2.0 **सॉफ्टवेयर का मेन विंडो**



	rank	score	rank	score	rank	score	rank	score	rank	score
Assigned sample	1	%	2	%	3	%	4	%	5	%
/Unk(MRT)	Marathwada	98.287	Jafrabadi	1.711	Murrah	0.002	Mehasana	0.000	Banni	0.000
/Unk(JFR)	Jafrabadi	99.995	Banni	0.005	Mehasana	0.000	Murrah	0.000	Marathwada	0.000
/Unk(Banni)	Banni	99.972	Jafrabadi	0.023	Mehasana	0.005	Marathwada	0.000	Murrah	0.000
/Unk(Meh)	Mehasana	92.083	Banni	4.858	Marathwada	2.913	Murrah	0.134	Jafrabadi	0.012
/Unk(Murrah)	Murrah	99.880	Jafrabadi	0.116	Marathwada	0.004	Mehasana	0.000	Banni	0.000

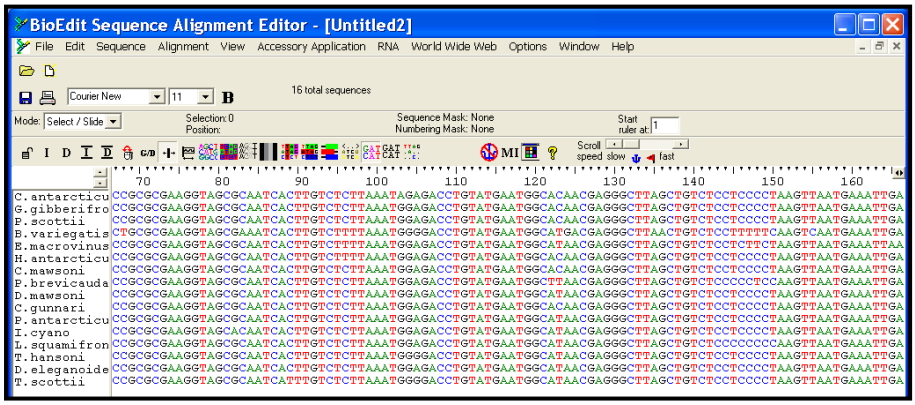
चित्र 2 : संदर्भ आंकड़ों सहित भैंस की 5 अज्ञात नस्लों की पहचान

BioEdit

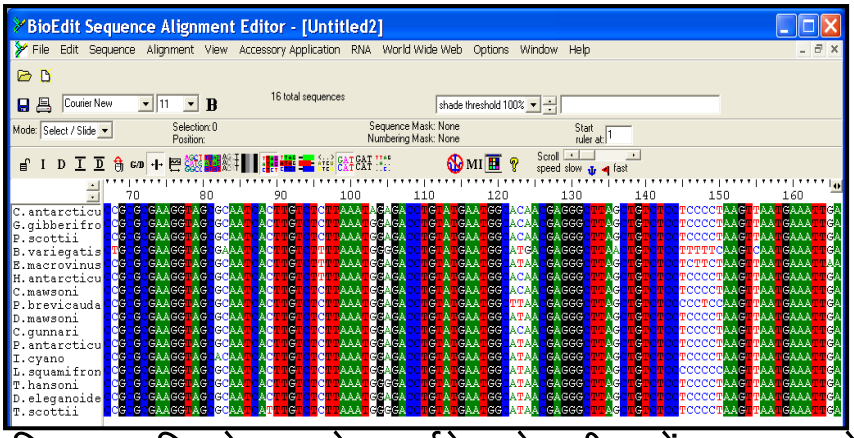
BioEdit माउस द्वारा चलने वाला उपयोग में सरल सीक्वेंस एलाइनमेंट एडिटर तथा क्रम विश्लेषण युक्ति है। इस युक्ति से सर्वाधिक सरल क्रम की साज संभाल, एलाइनमेंट की एडिटिंग तथा ऐसे अनेक कार्य किए जा सकते हैं जो अनुसंधानकर्ता दिन–प्रतिदिन करना चाहते हैं तथा इसके द्वारा कुछ मूल क्रम भी विश्लेषित किए जा सकते हैं। उदाहरण के लिए चित्र 1 और 2 में विभिन्न जीवाण्विक प्रभेदों के भिन्न न्यूक्लियोडाइड क्रम एलाइन किए गए हैं।

चरण निम्नानुसार हैं :

फाइलः → नया एलाइनमेंट → इम्पोर्ट →गौण अनुप्रयोग → क्लस्टल डब्ल्यू एलाइनमेंट→मल्टीपल एलाइनमेंट (चित्र 3) तथा एलाइनमेंट परिणामों को देखने के लिए : देखना → व्यू मोड → पहचान/समानता (चित्र 4)।



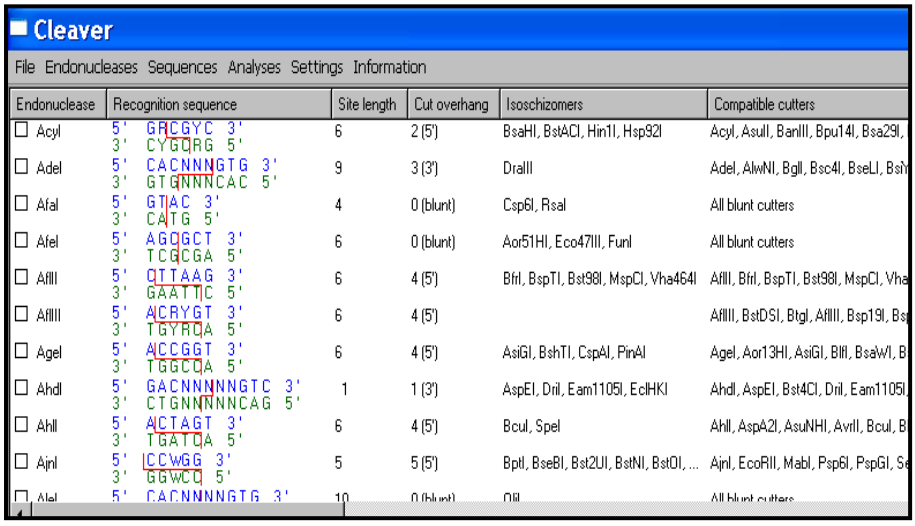
चित्र 3. न्यूक्लियोटाइड क्रम डेटा (16 विभिन्न सूक्ष्मजैविक प्रभेद)



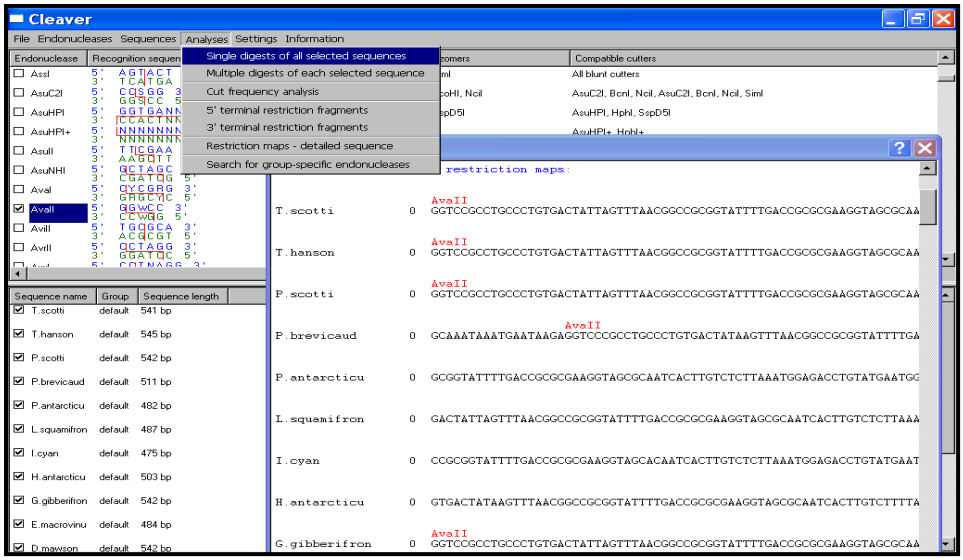
चित्र 4. न्यूक्लियोटाइड भेद दर्शाने वाले सभी क्रमों का एलाइनमेंट

क्लीवर

क्लीवर उन रिस्ट्रिक्शन एंडारे न्यूक्लियोएज पहचान स्थलों का अनुप्रयोग है जो कुछ टैक्सा में पाए जाते हैं (जारमैन, 2006)। टैक्सा के बीच डी.एन.ए. खण्ड रेस्ट्रिक्शन में भेद के पैटर्न वर्गीकरण विज्ञानी पहचान के लिए अनेक नैदानिक मूल्यांकनों का आधार हैं जिनका उपयोग मिश्रित स्रोतों से डी.एन.ए. के पूलों से कुछ टैक्सा के डी.एन.ए. को हटाने के लिए कुछ प्रक्रियाओं में किया जाता है। आर्थोलॉगस डी.एन.ए. क्रमों के समूहों के क्लीवर विश्लेषण रिस्ट्रिक्शन डाइजेशन से विभिन्न टैक्सा से व्युत्पन्न खण्डों के बीच रेस्ट्रिक्शन पैटर्न में मौजूद भेदों को पहचानना संभव होता है। यह क्लीवर वैबसाइट (http://cleaver.sourceforge.net/) पर निःशुल्क उपलब्ध है। यह कार्यक्रम कम्प्यूटरों के लिए एक स्क्रिप्ट के रूप में चलाया जा सकता है जिसमें फाइटन 2.3 हो तथा आवश्यक अतिरिक्त मॉड्यूल इंस्टाल किया गया हो। यह जीएनयू/लीनक्स, यूनिक्स, मैकओएसएक्स तथा विंडो प्लेटफार्मों पर चलाया जा सकता है। विंडोस तथा मैकओएसएक्स परिचालन प्रणालियों के लिए स्टेंडएलोन के कार्यान्वयनशील संस्करण भी उपलब्ध हैं। इस सॉफ्टवेयर को उपयोग करने का प्रोटोकाल चित्र 5 और चित्र 6 में दिखाया गया है।



चित्र 5. क्लीवर सॉफ्टवेयर का मुख्य विंडो

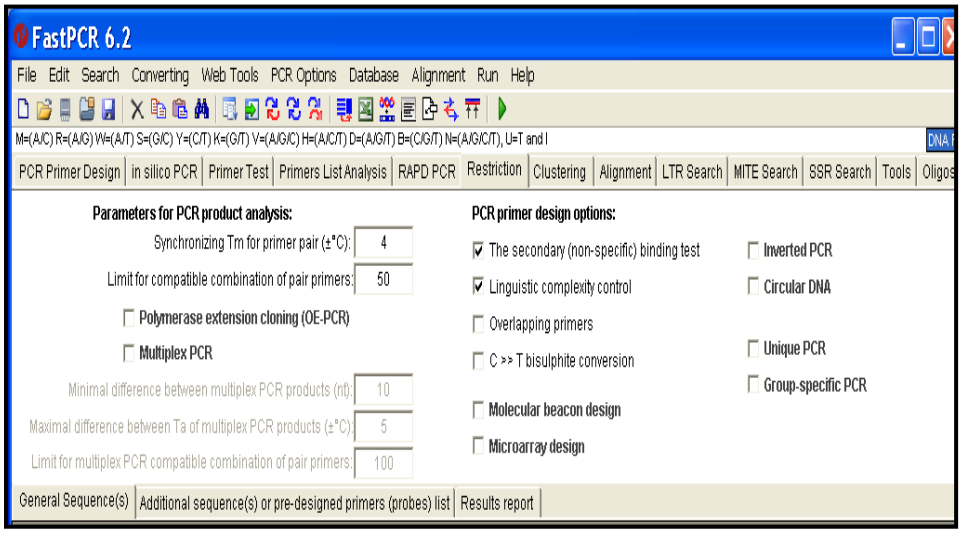


चित्र 6. क्लीवर सॉफ्टवेयर का उपयोग करके विभिन्न जीवाणिक जीनोमों के भिन्न क्रमों का रेस्ट्रिक्शन मानचित्र विश्लेषण

Fast PCR

Fast PCR पी.सी.आर. प्राइमरों या प्रॉब डिजाइन, *इन सिलिको* पी.सी.आर., ओलिगोन्यूक्लियोटाइड एसेम्बली और विश्लेषण, एलाइनमेंट और रिपीट सर्चिंग के लिए एक समेकित युक्ति है (चित्र 7)। यह सॉफ्टवेयर सभी युक्तियों के लिए सामान्य और डीजेनेरेटिड प्राइमरों का मिला–जुला उपयोग करता है तथा निकटतम ताप गतिकीय प्राचलों पर आधारित गलन तापमान की गणना के लिए भी इसका उपयोग किया जाता है। सम्पूर्ण जीनोम या गुणसूत्र की सूची के विरूद्ध *इन सिलिको* (वर्चुअल) पी.सी.आर. प्राइमर या प्रॉब सर्चिंग या *इन सिलिको* पी.सी.आर. की सूची में वहनीय पीसीआर उत्पादों का पूर्वानुमान, विशेष प्राइमरों या प्रॉबों की बेमेल स्थिति की खोज की क्षमता जैसे अनुप्रयोग शामिल हैं। वृहत प्राइमर परीक्षण जैसे मानक या डीजेनेरेट ऑलिगोन्यूक्लियोटाइडों के लिए गलनांक की गणना, प्राइमर पी.सी.आर. दक्षता, प्राइमरों की भाषा संबंधी जटिलता तथा डाइल्यूशन व रीसस्पेंशन कैल्कुलेटर को भी इसमें शामिल किया गया है। सभी प्राइमर द्वितीयक संरचनाओं के लिए प्राइमरों (प्रॉबों) का विश्लेषण किया जाता है जिसमें जी–क्वाड्रूपलैक्सिस पहचान, हेयरपिन, सैल्फ–डाइमर और क्रॉस डाइमर जो प्राइमर युग्म हैं, भी शामिल हैं। इस युक्ति में लंबे क्रमों को संभालने की क्षमता है साथ ही इसके द्वारा नाभिक अम्ल या प्रोटीन क्रमों का भी अनुरक्षण किया जा सकता है तथा इससे प्रत्येक दिए गए क्रम के लिए वैयक्तिक कार्य करने तथा प्राचलों को समझने के अलावा एक ही रन में अनेक विभिन्न कार्यों को मिल–जुलकर सम्पन्न करने की भी क्षमता है। इसके द्वारा सीक्वेंस एडिटिंग तथा डेटाबेस विश्लेषण भी किया जा सकता है। रिपीट्स के विभिन्न प्रकारों की कारगर तथा पूर्ण पहचान के लिए युक्तियां विकसित की गई हैं (डी.एन.ए. आधारित सिग्नेचर के लिए) तथा इनका उपयोग कार्यक्रम को दर्शाने हेतु भी किया गया है। कार्यक्रम में क्रम विश्लेषण के लिए विभिन्न जैव सूचना विज्ञानी युक्तियां शामिल हैं जिनमें जी.सी. या एटी स्क्यो, सी.जी. अंश तथा प्यूरीन–पाइरीमिडिन स्क्यू भाषायी क्रम जटिलता, जेनेरेशन रेंडम डी.एन.ए. क्रम, रेस्ट्रिक्शन विश्लेषण, क्रमों के क्लस्ट्रिंग की

सुविधा और कंसेन्सस क्रम सृजन, क्रमों की समानता तथा कंजर्वेन्सी विश्लेषण भी शामिल है।



चित्र 7. Fast PCR सॉफ्टवेयर का मुख्य विंडो

एस.एस.आर. सर्च या किसी अन्य विश्लेषण के लिए हमें केवल नोट पेड फाइल में डेटा फाइल तैयार करने तथा उसे मुख्य विंडो में इम्पोर्ट करने की आवश्यकता होती है। हम अपनी आवश्यकतानुसार मेन विंडो में ऑप्शन को देखते हुए रन/एस.एस.आर. सर्च/प्राइमर लिस्ट विश्लेषण पर क्लिक करके आंकड़ों को इम्पोर्ट करके उनका विश्लेषण कर सकते हैं।

संदर्भ

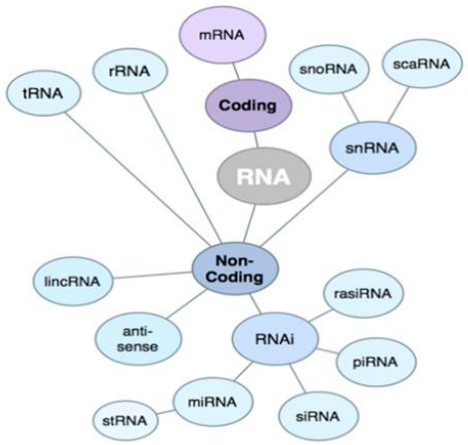
- बुस्तामांते, सीडी, निएल्सन, आर. और हार्टल, डीएल (2003). मैक्सिमम लाइकलीहुड एंड बेइसियन मैथड्स फॉर एस्टीमेटिंग द डिस्ट्रीब्यूशन ऑफ सलेक्टिव इफेक्ट्स अमंग क्लासिस और म्यूटेशंस यूजिंग डी.एन.ए. पॉलीमॉर्फिज्म डेटा. थ्योरिटिकल पापुलेशन बायोलॉजी 63 : 91–103.
- कोरेंडर, जे., वांडमैन, पी. और सिलेन्पा, एमजे. (2003) बायेसियन एनालिसिस ऑफ जेनेटिक डिफ्रंसिएशन बिटविन पोपुलेशन. जेनेटिक्स. 163 : 367–374.
- गोल्डस्टेइन, डीबी, लिनेर्स, एआर, कैवाली–स्फोर्जा, एलएल और फेल्डमैन, एमडब्ल्यू (1995). जेनेटिक एब्सुलूट डेटिंग बेस्ड ऑन माइक्रोसेटेलाइट एन ओरिजन ऑफ मॉडर्न ह्यूमंस. पीएनएएस यूएसए. **92**: 6723–6727.
- हैबर्ट, पीडीएन, पैंटोन, ईएच, बर्न्स, जेएम, जांजेन, डीएच. और हालवाच्स, डब्ल्यू. (2004). टैन स्पीसीज इन वन : डी.एन.ए. बारकोडिंग रिवील्स क्रिप्टिक स्पीसीज इन द न्यूट्रोपिकल स्कीपर बटरफ्लाई एस्ट्राप्टेस्फलगरेटर. प्रोस. नेशनल अकाडमी ऑफ साइंस, यू.एस.ए. **101(41)** : 14812–14817

- हैबर्ट, पीडीएन, स्टोइकल, एमवाई, जैमलैक, टीएस और फ्रांसिस, सीएम. (2004b). आइडेंटीफिकेशन ऑफ बर्डस थ्रू डी.एन.ए. बारकोड्स. पी.एल.ओ.एस. बायोल. **2(10)**: 1657–1663.
- जार्मेन, (2006). क्लीवर : सॉफ्टवेयर फॉर आइडेंटीफाइंग टैक्सॉन स्पेसिफिक रेस्ट्रिक्शन एंडो न्यूक्लिएज रिकग्नीशन साइटस. बायोइंफोमेर्टिक एडवांस एक्सिस (http://bioinformatics.oxfordjournals.org/content/early/2006/06/20/bioinformatics.btl339.full.pdf.½
- क्रैस. डब्ल्यूजे, वुर्डेक केजे, जीमेर, ई.ए., वेट, एलए और जांजेन, डीएच (2005). यूज ऑफ डी.एन.ए. बारकोड्स टु आईडेंटीफाई फ्लावरिंग प्लांट्स, प्रोसीडिंगस नेशनल एकाडमी ऑफ साइंस, यूएसए, **102(23) :** 8369–8374.
- पीटकाउ, डी., काल्वर्ट, डब्ल्यू, स्टर्लिंग, आई और स्ट्रोबैक, सी. (1995). माइक्रोसेटेलाइट एनालिसिस ऑफ पोपुलेशन स्ट्रक्चर इन कैनिडियन पोलर बियर्स. मॉलीक्यूलर इकोलॉजी. **4** : 347–354.
- रानाल्ला, बी. और माउंटेन, जेएल., (1997). डिटेक्टिंग इमीग्रेशन बाइ यूजिंग मल्टी लॉकस जीनोटाइप्स. पी.एन.ए.एस., यू.एस.ए. **94** : 9197–9221.
- सासाजाकी, एस., होसोकावा, डी., इशीहारा, आर., एइहारा, एच., ओयामा के., मैनन, एच. (2011). डेवलपमेंट ऑफ डिस्क्रिमिनेशन मार्कर्स बिटवीन जापानीस डोमेस्टिक एंड इम्पोर्टिड बीफ. एनिमल साइंस जर्नल. **82 (1)** : 67–72.
- सुएकावा, वाई., एइहारा, एच., अराकी, एम., होसोकावा, डी., मैन्नन, एच. सासाजाकी, एस. (2010). डेवलपमेंट ऑफ ब्रीड आइडेंटीफिकेशन मार्कर्स बेस्ड ऑन ए बोवाइन 50के एस.एन.पी. ऐरे. मीट साइंस. **85 (2)** : 285–288.

आर.एन.ए. सेक डेटा विश्लेषण

भूमिका

आगामी-पीढ़ी अनुक्रमण/ नेक्स्ट-जेनेरेशन सीक्वेंसिंग (एन.जी.एस.) तकनीक के आगमन ने जीनोमिक अध्ययन को बदल दिया है। एन.जी.एस. तकनीक का एक महत्वपूर्ण अनुप्रयोग ट्रांसक्रिप्टॉम का अध्ययन है। कोशिका में सभी आर.एन.ए. (RNA) अणुओं के पूर्ण संग्रह को ट्रांसक्रिप्टॉम कहा जाता है। विभिन्न प्रकार के आर.एन.ए. जिन्हें अब तक वर्गीकृत किया गया है, उन्हें चित्र 1 में दिखाया गया है। इन सभी अणुओं को ट्रांसक्रिप्टॉम कहा जाता है क्योंकि वे ट्रांसक्रिप्सन की प्रक्रिया द्वारा निर्मित होते हैं।



चित्र 1. विभिन्न प्रकार के आर.एन.ए.

एम.आर.एन.ए (mRNA) की एन.जी.एस. अर्थात् आर.एन.ए.-सेक (RNA-Seq) जैविक प्रयोगों में जीन अभिव्यक्ति (एक्सप्रेसन) को मापने के लिए एक मानक बन गया है। कम्प्यूटेशनल और सांख्यिकीय दृष्टिकोण से आर.एन.ए.-सेक डेटा (आंकड़ा) विश्लेषण सबसे संभावित शोध क्षेत्र रहा है जो ट्रांसक्रिप्टोमिक स्तर पर जीन की भूमिकाओं में एक अंतर्दृष्टि प्रदान कर सकता है। आर.एन.ए.-सेक डेटा उत्पन्न करने के लिए कई मशीनें/ प्रोटोकॉल उपलब्ध हैं, जैसे, इलुमिना (मिसेक, नेक्स्टसेक, हायसेक, नोवासेक), आयन टोरेंट (प्रोटॉन, पर्सनल जीनोम मशीन), सोलिड, रोच 454, आदि। आर.एन.ए.-सेक आंकड़ों का विश्लेषण माइक्रोएरे डेटा विश्लेषण से विभिन्न पहलुओं में भिन्न होता है जैसे डेटा की प्रकृति, सामान्यीकरण के तरीके और डिफ्रेंसियल एक्सप्रेसन विश्लेषण।

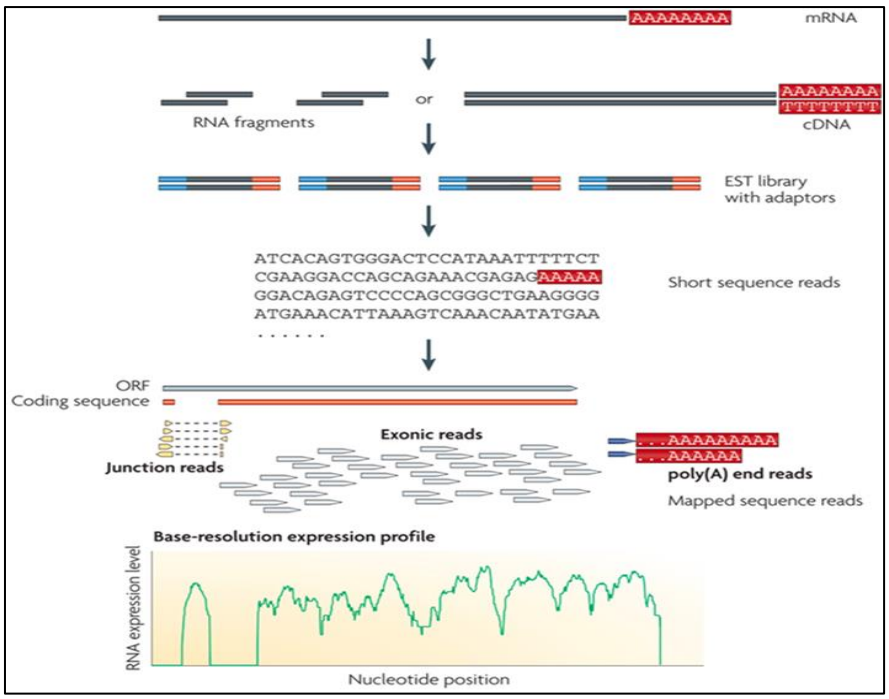
आर.एन.ए.-सेक के कई अनुप्रयोग हैं: ट्रांसक्रिप्टॉम/ आर.एन.ए. एक्सप्रेसन के स्तरों का परिमाणन, नई जीन की खोज, जीन एनोटेशन, विभिन्न स्थितियों के बीच डिफ्रेंसियली एबंडेंट/ एक्सप्रेस्ड फीचर्स (जीन्स/ ट्रांसक्रिप्ट्स/ एक्सॉन्स) का पता लगाना, स्प्लाइसिंग घटनाओं का पता लगाना, इंट्रॉन्स और एक्सॉन्स की सीमाओं की पहचान, इत्यादि।

आर.एन.ए.-सेक प्रयोग

आर.एन.ए.-सेक प्रयोग में कई महत्वपूर्ण चरण हैं: 1. डेटा उत्पादन (प्रयोगात्मक डिजाइन, सैम्पल संग्रह, सीक्वेंसिंग डिजाइन और गुणवत्ता नियंत्रण), 2. एक्सप्रेसन वेल्युस प्राप्त करने के लिए रीड्स (मैपिंग या एलाइनमेंट) का परिमाणन, 3. सामान्यीकरण; 4. डिफ्रेंसियल एक्सप्रेसन विश्लेषण। एक आम आर.एन.ए.-सेक प्रयोग को सारांशित करने के लिए मूल चरण निम्नानुसार हैं (कृपया चित्र 2 देखें):

- पहले शुद्ध आर.एन.ए. को सी.डी.एन.ए.(cDNA) में बदल दिया जाता है। फिर सीक्वेंसिंग लाइब्रेरी तैयार किया जाता है और एक एन.जी.एस. प्लेटफॉर्म पर सीक्वेंसिंग किया जाता है।
- सी.डी.एन.ए. अंशों के एक छोर (सिंगल-इंड) या दोनों छोर (पेयर्ड-इंड) से लाखों लघु सीक्वेंसिंग रीड्स उत्पन्न होते हैं।
- इन सीक्वेंस की मैपिंग संदर्भ (रिफरेंस) जीनोम से की जाती है।
- जाने हुए (ज्ञात) फीचर्स के लिए मैप की गई रीड की संख्या (रीड काउंट्स) को एक तालिका में दर्ज और संक्षेपित किया जाता है।

फीचर्स या तो जीन, ट्रांसक्रिप्ट (या अल्टरनेटिव ट्रांसक्रिप्ट), एलील स्पेसिफिक एक्सप्रेसन या एक्सॉन लेवल एक्सप्रेसन पर हो सकती हैं। उदाहरण के लिए, यदि F फीचर्स और N सैम्पल हैं, तो रीड काउंट्स की एक तालिका गैर- निगेटिव पूर्णांक का $F \times N$ मैट्रिक्स है।



चित्र 2. सामान्य आर.एन.ए.-सेक प्रयोग

आर.एन.ए.-सेक डेटा विश्लेषण

आर.एन.ए.-सेक विश्लेषण के लिए उपलब्ध कुछ ओपन सोर्स सॉफ्टवेअर इस प्रकार हैं:

रौ रीड डेटा (FASTQ फाइल्स) की गुणवत्ता जांच

- फास्ट क्यू. सी. (FastQC), एन.जी.एस.क्यू.सी. (NGSQC)

डेटा प्रीप्रोसेसिंग

- फास्टएक्स टूलकिट (FASTX toolkit), शॉर्टरीड (ShortRead), ट्रिम्मोमैटिक (Trimmomatic), सैमटूल्स (Samtools)

शॉर्ट रीड्स एलाइनर्स (Short reads aligners)

- बोटाई (Bowtie), टॉपहैट (TOPHAT), बी.डब्ल्यू.ए. (BWA), नोवोएलाइन (Novoalign), स्टार (STAR), आदि

***डी नोवो* अस्सेम्ब्लेर्स (*de novo* assemblers)**

- सोपडीनोवो-ट्रांस (SOAPdenovo-Trans), ट्रांस-अबिस (Trans-AbySS), ट्रिनिटी (Trinity), स्पेड्स (SPAdes)

फीचर परिमाणन

- रौ रीड काउंट डेटा: एच.टी.सेक-काउंट (htseq-count), फीचरकाउंट्स (featureCounts)
- एक्सप्रेसन वेल्युस की परिमाणन करने के अन्य तरीके: कफ़लिंक्स (Cufflinks), स्ट्रिंगटाई (Stringtie), आर.एस.ई.एम. (RSEM), सेलफ़िश (Sailfish)

एक्सप्रेसन अध्ययन

- कफ़लिंक्स पैकेज (Cufflinks package)
- आर पैकेजेस (R packages): डी.ई.सेक (DESeq), डी.ई.सेक2 (DESeq2), एज.आर (edgeR), आदि

विज़ुअलाइज़ेसन

- कमेआरबंड (CummeRbund), आई.जी.वी. (IGV), बेडटूल्स (Bedtools), यू.सी.एस.सी. जीनोम ब्राउज़र (UCSC Genome Browser), आदि

आर.एन.ए.-सेक रीड काउंट डेटा का एक उदाहरण

एक विशिष्ट आर.एन.ए.-सेक प्रयोग में, सैम्पल्स की सीक्वेंसिंग की जाती है और रीड्स की मैपिंग रिफरेंस जीनोम से की जाती है। प्रत्येक रिफरेंस जीन्स से मैप किए गए रीड्स की संख्या (रीड काउंट्स) की गणना की जाती है। मान लीजिये एक RNA-Seq प्रयोग में N सैम्पल्स हैं। आगे मान लीजिये कि स्थिति/ समूह C_i $(i = 1, 2)$ में j^{th} सैम्पल $(j = 1, 2, \dots, n_i)$ के जीन G_k $(k = 1, 2, \dots, K)$ में मैप किए गए रीड्स की संख्या Y_{ijk} है। आमतौर पर काउंट डेटा (Y_{ijk}) की मोडलिंग प्वाइजन डिस्ट्रिब्युसन या निगेटिव बायनोमियल डिस्ट्रिब्युसन द्वारा किया जाता है। एक काल्पनिक केस-कंट्रोल अध्ययन के लिए रीड काउंट की एक तालिका नीचे दी गई है (चित्र 3)।

	Conditions/ Treatment groups											
	C_1(Case)						C_2(Control)					
Samples → / Genes ↓	$S_{1,1}$	$S_{1,2}$	...	$S_{1,j}$	...	S_{1,n_1}	$S_{2,1}$	$S_{2,2}$	...	$S_{2,j}$	...	S_{2,n_2}
G_1	21	30	...	25	...	5	65	61	...	52	...	25
G_2	0	3	...	1	...	0	7	2	...	0	...	6
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
G_k	198	122	...	162	...	51	302	245	...	102	...	29
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
G_K	2	1	...	0	...	1	1	0	...	0	...	1

चित्र 3. एक काल्पनिक केस-कंट्रोल अध्ययन के लिए रीड काउंट की तालिका

डिफ्रेंसियल एक्सप्रेसन विश्लेषण के लिए विभिन्न आर पैकेजेस (R packages) उपलब्ध हैं जैसे कि एज.आर (edgeR), डी.ई.सेक (DESeq), डी.ई.सेक2 (DESeq2), आदि। डिफ्रेंसियल एक्सप्रेसन विश्लेषण करने से पहले सामान्यीकरण की आवश्यकता होती है। सामान्यीकरण के अनेक तरीके हैं जैसे कि आर.पी.के.एम. (रीड्स अलाइन्ड पर किलोबेस ऑफ एक्सॉन पर मिलियन रीड्स मैप्ड), एफ.पी.के.एम. (फ्रेगमेंट्स अलाइन्ड पर किलोबेस ऑफ एक्सॉन पर मिलियन फ्रेगमेंट्स मैप्ड) और टी.पी.एम. (ट्रांसक्रिप्ट्स पर किलोबेस मिलियन)। ये तरीके डेटा के सामान्यीकरण के लिए जीन की लंबाई और सीक्वेंसिंग डेप्थ का उपयोग करते हैं।

निष्कर्ष

आर.एन.ए.-सेक अभी भी उपयोग में है और पहले से विकसित ट्रांसक्रिप्टॉमिक विधियों पर इसके लाभ स्पष्ट हैं। मगर आर.एन.ए.-सेक प्रयोगों के उपयोग से जुड़ी अनेक चुनौतियां हैं जैसे लाइब्रेरी का निर्माण, जैव सूचना विज्ञान की समस्या (बड़े डेटा सेट का संचयन, पुनर्प्राप्ति और प्रोसेसिंग; मैपिंग और असेंबली की समस्या), सीक्वेंस/ ट्रांसक्रिप्टॉम कवरेज बनाम लागत, ट्रांसक्रिप्टॉमिक विश्लेषण (इंट्रॉन्स और एक्सॉन्स की सीमाओं की पहचान के साथ-साथ नई जीन की खोज के लिए जीन मैपिंग; स्प्लाइसिंग घटनाओं का पता लगाना; जटिल प्रयोगों में जीन एक्सप्रेसन का अध्ययन करने के लिए ट्रांसक्रिपटॉम/ आर.एन.ए. एक्सप्रेसन के स्तरों का परिमाणन), इत्यादि। आने वाले भविष्य में यह मौजूदा तकनीक पर अधिक सुधार के साथ और बेहतर हो जाएगा तथा अन्य अनुप्रयोगों के लिए यह माइक्रोएरे जैसी तकनीक की जगह ले लेगा।

संदर्भ

- वांग जेड., गेरस्टीन एम., स्नाइडर एम. (2009). आर.एन.ए.-सेक: ट्रांसक्रिप्टॉमिक्स के लिए एक क्रांतिकारी टूल, *नेचर रीव्यू जेनेटिक्स*, 10 (1), 57-63।
- एंडर्स एस., ह्यूबर डब्ल्यू. (2010). सीक्वेंस काउंट डेटा के लिए डिफ्रेंसियल एक्सप्रेसन विश्लेषण,*जीनोम बायोलोजी*, 11, R106।
- मोर्टाज़ावी ए., विलियम्स बी.ए., मैक्यू के., शेफ़र एल., और वोल्ड बी. (2008). आर.एन.ए.-सेक द्वारा मैमलियन ट्रांसक्रिप्टॉम्स का मैपिंग और परिमाणित करना, *नेचर मेथड्स*, 5 (7), 621-628।
- शेंड्योर जे., जी एच. (2008). नेक्स्ट-जेनेरेशन आर.एन.ए. सीक्वेंसिंग, *नेचर बायोटेक्नोलॉजी*, 26, 2514-2521।
- ब्रायन जे. एच. और माइकल सी. जेड. (2010). आर.एन.ए.-सेक विश्लेषण को आगे बढ़ाना, *नेचर बायोटेक्नोलॉजी*, 28, 421-423।

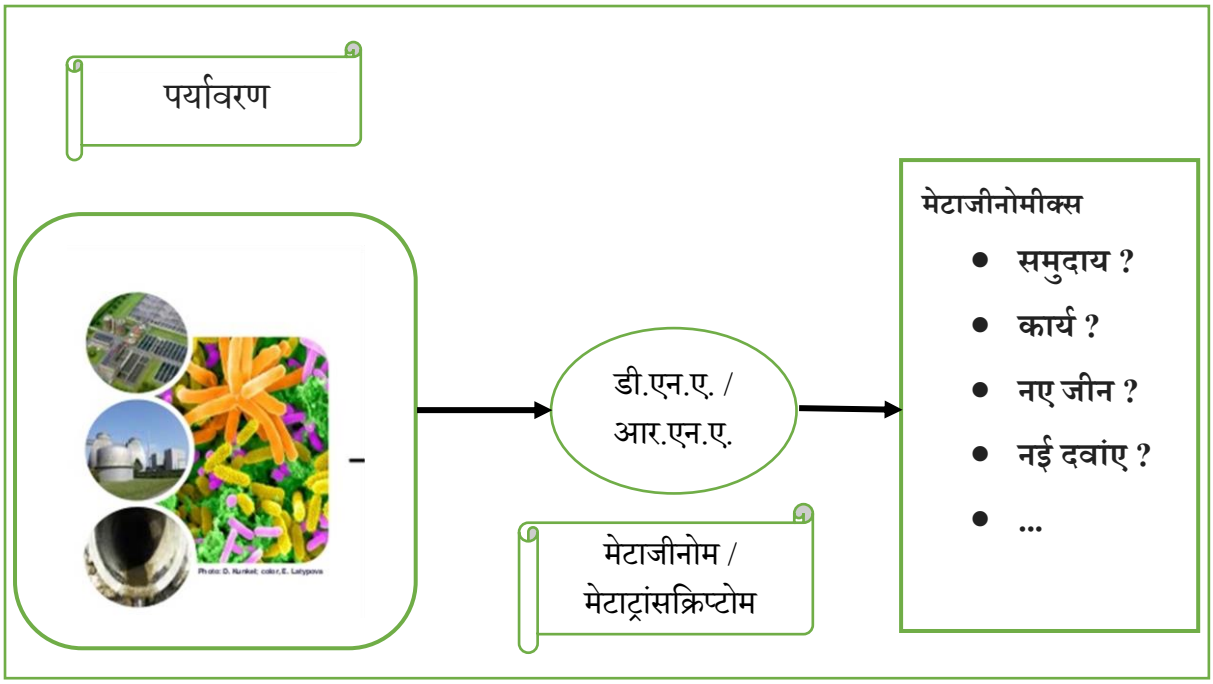
कृषि में मेटाजीनोमिक्स की भूमिका

1. परिचय

जीवमंडल में सूक्ष्मजीवों (व्हिटमैन आदि, 1998) का वर्चस्व है, जो कि मेडिसिन, इंजीनियरिंग और कृषि में प्रायोगिक महत्व रखते हैं (स्लोअन आदि, 2006)| उनके इसी महत्व के कारण, सूक्ष्मजीवों की आनुवांशिक और जैविक विविधता वैज्ञानिक अनुसंधान का एक महत्वपूर्ण क्षेत्र है| सूक्ष्मजीवों के महत्व की स्पष्टता के बावजूद, उनकी विविधता के बारे में बहुत कम जानकारी है, उदाहरण के लिए पर्यावरण में कितनी प्रजातियां मौजूद हैं, अथवा प्रत्येक प्रजाति क्या करती है और इसके पारिस्थितिक कार्यों की जानकारी (सिंह आदि, 2008)। अभी तक, सूक्ष्मजीवों के संवर्धन में हुई सीमाओं के कारण इन महत्वपूर्ण सवालों के जवाब देने के लिए कोई उपयुक्त तकनीक उपलब्ध नहीं थी। पारंपरिक तरीकों से उन्हीं सूक्ष्मजीवों का संवर्धन किया जा सकता है जो प्रायोगिक स्थितियों में बढ़ते हैं| हालांकि, यह व्यापक रूप से स्वीकार किया जाता है कि पर्यावरण में 99% तक सूक्ष्मजीवों को आसानी से कल्चर नहीं किया जा सकता।

इस प्रकार, जैव प्रौद्योगिकी के लिए अधिकांश रोगाणुओं का वर्णन और मूल्यांकन नहीं किया गया है। इस समस्या के समाधान के लिए अलग प्रकार की डी.एन.ए. आधारित आणविक तरीकों को विकसित किया गया है | इन विधियों ने माइक्रोबियल विविधता और पारिस्थितिकी की हमारी समझ को काफी प्रभावित किया है (डीलांग और कार्ल, 2005)| सामान्य तौर पर, 16 एस. आर.एन.ए. (16S rRNA) जीन विश्लेषण पर आधारित विधियों ने पर्यावरण और वातावरण में मौजूद टैक्सा और प्रजातियों के बारे में विस्तृत जानकारी प्रदान की है। हालाँकि, ये आंकड़े आम तौर पर समुदाय के भीतर विभिन्न रोगाणुओं की कार्यात्मक भूमिका के बारे में बहुत कम जानकारी प्रदान करते हैं (स्ट्रीट और स्मिज़, 2004)| 1990 के दशक के बाद, इन कठिनाइयों पर काबू पाने के लिए एक नई तकनीक शुरू की गई है जिसे ‘मेटाजीनोमिक्स’ कहा जाता है|

"मेटाजीनोम" शब्द का इस्तेमाल पहली बार जे. हैंडेल्समैन द्वारा 1998 में किया गया था जो एक पर्यावरण नमूने में आनुवंशिक सामग्री की कुल राशि के वर्णन के लिए था| मेटाजीनोमिक्स आणविक जीव विज्ञान और आनुवंशिकी को इस प्रकार जोड़ती है जिससे कि (पर्यावरण) नमूनों से आनुवांशिक सामग्री को पहचाना जा सके और उनका वर्णन किया जा सके (चित्र 1)| मेटाजीनोमिक्स एक ऐसा उभरता हुआ क्षेत्र है जिसमें सूक्ष्मजीवों के पूरे समुदाय पर जीनोमिक विश्लेषण कर सूक्ष्मजीव प्रजातियों को प्रयोगशाला में कल्चर करे बिना अलग-अलग किया जा सकता है (पैट्रिक आदि, 2005) |



चित्र 1: मेटाजीनोमीक्स का अवलोकन

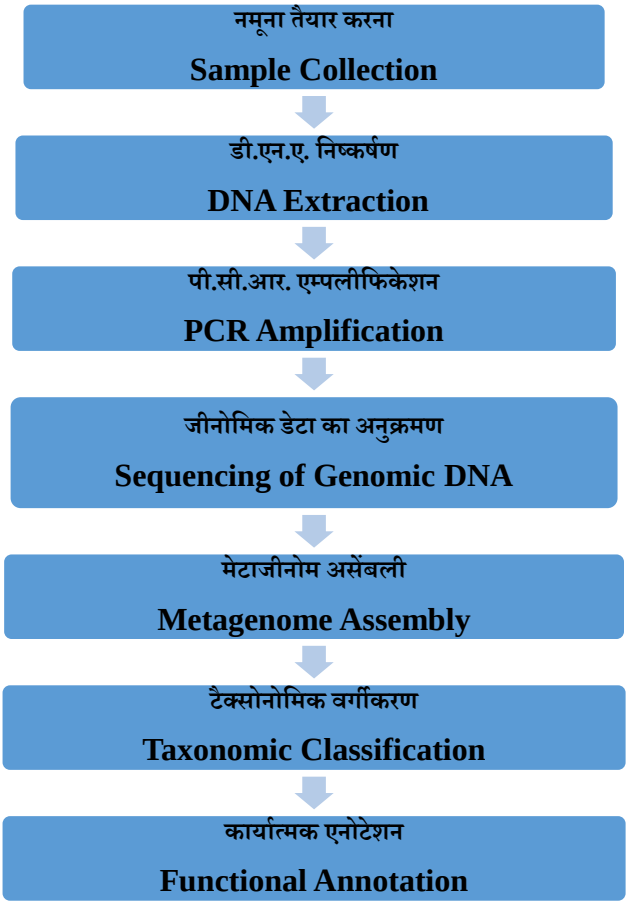
2. मेटाजीनोमीक्स में बुनियादी कम्प्यूटेशनल कार्य

मेटाजीनोमीक्स में तीन प्रकार के बुनियादी कार्य संगणनात्मक विधियों के द्वारा किए जाते हैं

- टैक्सोनोमिक विश्लेषण(कौन वहाँ है?)
- कार्यात्मक विश्लेषण (वे क्या कर रहे हैं?)
- तुलनात्मक विश्लेषण (वे कैसे तुलना करते हैं?)

3. मेटाजेनोमिक्स विश्लेषण पाइपलाइन

मेटाजीनोमिक्स डेटा के विश्लेषण के लिए एक सामान्य पाइपलाइन में मूलतः सात चरण होते है जो इस प्रकार है (1) नमूना तैयार करना, (2) डी.एन.ए. निष्कर्षण, (3) पी.सी.आर. विस्तारण (Amplification), (4) जीनोमिक डेटा का अनुक्रमण, (5) मेटाजीनोम असेंबली, (6) टैक्सोनोमिक वर्गीकरण और (7) कार्यात्मक एनोटेशन| यह चित्र 2 में भी दिखाया गया है।



चित्र 2: माइक्रोबियल विविधता विश्लेषण पाइपलाइन

4.माइक्रोबियल विविधता विश्लेषण के लिए सॉफ्टवेयर उपकरण

मेटाजीनोम के अनुक्रमण के बाद अनेक कार्यों के लिए सॉफ़्टवेयर टूल्स का प्रयोग किया जाता है जैसे मेटाजीनोम की असेंबली, मेटाजीनोम का टैक्सोनोमिक वर्गीकरण और इसका कार्यात्मक एनोटेशन इत्यादि। यह खंड इन कार्यों को करने के लिए प्रयुक्त किए जाने वाले सॉफ्टवेयर उपकरण का वर्णन करता है।

4.1 मेटाजीनोम असेंबली टूल्स

• **मेटा आई.डी.बी.ए. (MetaIDBA)** – यह एक डी ब्रून (de bruijn) ग्राफ पर आधारित असेंबलर है जिसमें कई असेंबलर का संग्रह हैं एवं हर असेम्बलर एक विशिष्ट कार्य करता हैं। मेटाजीनोम की असेंबली मेटा आई.डी.बी.ए. द्वारा की जाती है (पेंग आदि, 2011)।

- **मेटा वेल्वेट (Meta Velvet)** – यह एक डी ब्रून (de bruijn) ग्राफ असेंबलर वेल्वेट पर आधारित है (नामिकी आदि, 2012; ज़र्बिनो आदि, 2008)| इस असेंबलर में के-मर इंडेक्सिंग और ग्राफ बिल्डिंग के लिए वेल्वेट के velveth और velvetg नामक मॉड्यूल को आधार बनाकर असेंबली किया गया है (ज़र्बिनो आदि, 2008))|

- **ओमेगा (Omega) (हैदर आदि, 2014)** – यह असेंबलर ओवरलैप (Overlap) आधारित स्ट्रिंग ग्राफ़ विधि (String Graph Approach) (मायर्स, 2005) का उपयोग करता है| आमतौर पर यह लंबे अनुक्रमण रीड (Long Reads) डेटा की असेंबली के लिए उपयोग किया जाता है। यह उपकरण मेटाजीनोम के इल्युमिना (illumina) अनुक्रमण डेटा के लिए डिज़ाइन किया गया है।

- **मेगाहिट (MegaHit) (ली. आदि, 2015)** "सक्सिंट डी ब्रून (succinct *de Bruijn)*" नामक एक नई डेटा संरचना का उपयोग करता है| इस विधि में कम कंप्यूटर मेमोरी (memory requirements) की आवश्यकता होती है । मेगाहिट सिंगल (single) एवं युग्मित-अंत (paired end) रीड (read) के संकुचित (compressed) और असंकुचित (uncompressed) फ़ास्टा (fasta) और फ़ास्टक्यू (Fastq) फॉरमेट को स्वीकार करता है | यह स्टैण्डर्ड इनपुट द्वारा पाइपिंग इनपुट डेटा को भी स्वीकार करता है |

4.2 मेटाजीनोमिक डेटा के टैक्सोनोमिक वर्गीकरण के लिए सॉफ्टवेयर उपकरण

मेटाजीनोमिक डेटासे के टैक्सोनोमिक वर्गीकरण के लिए उपलब्ध सॉफ्टवेयर उपकरण मुख्यत तीन दृष्टिकोणों पर आधारित है| कुछ सॉफ्टवेयर होमोलॉजी पर, कुछ संरचना पर एवं कुछ प्रचुरता अनुमान पर आधारित है| इन दृष्टिकोणों के आधार पर कई टूल और सॉफ्टवेयर विकसित किए गए हैं जिनका विवरण इस खंड में दिया गया है।

- **मेगन (MEGAN)** (हसन आदि, 2011) - यह मेटाजेनोमिक डेटासेट के विश्लेषण के लिए एक ग्राफिकल इंटरफ़ेस है जो कि मेटाजेनोमिक डेटा के विश्लेषण के लिए होमोलॉजी आधारित विधि का प्रयोग करता है| मेटाजेनोमिक डेटा को पहले ब्लास्ट (BLAST) या डायमंड आदि सॉफ्टवेयर का इस्तेमाल करके जीनोम डेटाबेस से संरेखित (align) किया जाता है।

- **एम.जी. रास्ट (MG-RAST)** (विल्के आदि, 2015) – यह एक वेबसर्वर है जो मेटाजेनोमिक डेटासेट के टैक्सोनोमिक और कार्यात्मक विश्लेषण के लिए एक पाइपलाइन का प्रयोग करता है| कार्यात्मक और टैक्सोनोमिक असाइनमेंट के लिए डेटा को न्यूक्लियोटाइड और प्रोटीन डेटाबेस से संरेखित (Align) किया जाता हैं। यह फाइलोजेनेटिक (Phylogenetic) और कार्यात्मक सारांश भी दर्शाता हैं| इसमें तुलनात्मक मेटाजीनोमिक्स की सुविधा भी उपलब्ध है।

• **क्राकेन (KRAKEN)** (वुड और सल्ज्बेर्ग, 2014) – यह एक प्रमुख और तेज वर्गीकारक है जो के-मर (k-mer) पर आधारित होमोलॉजी विधि का उपयोग करता है।

• **फिम (Phymm)** (ब्रैडी और सल्ज्बेर्ग, 2009) – यह पहले सटीक और तेज़ मेटाजीनोमिक वर्गीकारकों में से एक था। इसने एक अंतर्वेशित मार्कोव (Interpolated Markov) मॉडल को 539 पूर्ण क्यूरेटेड जीनोम डेटाबेस पर प्रशिक्षित किया गया था| इस मॉडल को बहुत छोटे अनुक्रमों के वर्गीकरण के लिए इस्तेमाल किया जा सकता है|

• **रीटा (RITA-Rapid Identification of Taxonomic Assignments)** यह एक होमोलॉजी और संरचना पर आधारित सॉफ्टवेयर है जो बहुत छोटे यानि 50 बेस पेयर (bp) तक के अनुक्रमों को निर्दिष्ट कर सकता है (मैक डोनाल्ड आदि, 2012)|

4.3 मेटाजेनोमिक्स डेटा के एनोटेशन के लिए सॉफ्टवेयर उपकरण

मेटाजेनोमिक्स डेटा का एनोटेशन दो तरीकों से किया जा सकता है। पहली विधि से बड़े कंटिग (contigs) का एनोटेशन किया जा सकता है| दूसरी विधि द्वारा असंकलित रीड (unassembled reads) और छोटे रीड (reads) का एनोटेशन किया जा सकता है। एनोटेशन करते समय सबसे पहले विशेषताओं (features) को पहचाना जाता है और उसके बाद कल्पित जीन फ़ंक्शन और टैक्सोनोमिक पड़ोसियों की पहचान की जाती है| **3** मेटाजेनोमिक्स डेटा के एनोटेशन के लिए सॉफ्टवेयर उपकरण निम्नलिखित है|

- फ्राग जीन स्कैन (FragGeneScan)
- मेटाजीन एनोटेटर (Metagene Annotator)
- टी. आर. एन. ए. स्कैन एस. ई. (tRNAscanSE)
- क्रिस्प आर (CRISPR)
- केग (KEGG)
- कॉग (COG)

5. निष्कर्ष

इस लेख में विभिन्न माइक्रोबियल समुदायों के अनुक्रमो की असेंबली टैक्सोनोमिक वर्गीकरण और , कार्यात्मक एनोटेशन में प्रयुक्त होने वाले सॉफ्टवेयर उपकरणों की विस्तृत रूप से चर्च कीगई है। मेटाजीनोमिक्स डेटा की बढ़ती मात्रा के कारण बहुत तरह के बढ़िया सॉफ्टवेयर उपकरण की आवश्यकता है। ये स्वचालित उपकरण मेटाजेनोमिक्स डेटा में संग्रहीत ज्ञान को खनन में बहुत उपयोगी हैं।

संदर्भ:

- व्हिटमैन, डब्लू. बी., कोलमन, डी. सी, विबे, डब्ल्यू.जे.(1998). Prokaryotes: the unseen majority. *Proc Natl Acad Sci USA*, खंड 9, पृष्ठ संख्या 6578-6583.
- स्लोअन, डब्ल्यू.टी., लून, एम., वुडकॉक. एस., हेड, आई.एम., नीस, एस. और कर्टिस, टी.पी. (2006). Quantifying the role of immigration and chance in shaping prokaryote communit structure ,पर्यावरण .माइक्रोबोल., खंड 8, पृष्ठ संख्या 732-740.
- सिंह, बी, गौतम, एस.के., वी। वर्मा, एम .के. और सिंह.बी ,(2008). Metagenomics in animal gastrointestinal ecosystem: Potential biotechnological prospects ,Anaerobe (एनारोब.), खंड 14, पृष्ठ संख्या 138.
- व्हिटमैन, डब्लू. बी., कोलमन, डी. सी, विबे, डब्ल्यू.जे. (1998). Prokaryotes: the unseen majority. Proc Natl Acad Sci USA, खंड 9, पृष्ठ संख्या 6578-6583.
- स्लोअन, डब्ल्यू.टी., लून, एम., वुडकॉक. एस., हेड, आई.एम., नीस, एस. और कर्टिस, टी.पी. (2006). Quantifying the role of immigration and chance in shaping prokaryote community structure ,पर्यावरण .माइक्रोबोल., खंड 8, पृष्ठ संख्या 732-740.
- सिंह, बी, गौतम, एस.के., वी। वर्मा, एम .के. और सिंह .बी ,(2008). Metagenomics in animal gastrointestinal ecosystem: Potential biotechnological prospects ,Anaerobe (एनारोब.), खंड 14, पृष्ठ संख्या 138
- स्ट्रेइट, डब्ल्यू.आर ., श्मिटज़, आर).ए. .(2004Metagenomics – the key to the uncultured microbes. Curr. Opin. Microbiol. ,खंड 7, पृष्ठ संख्या 498-492
- पैट्रिक डी, श्लॉस, हैंडल्समैन .(2005) .जे ,Metagenomics for studying unculturable microorganisms: cutting the Gordian knot ,(जीनोम बीओएल) .Genome Biol ,खंड ,6पृष्ठ संख्या 229
- हंडेल्स्मेन, जे. (2004). मटाजेनोमिक्स: एप्लिकेशन ऑफ जेनोमिक्स टू अनकलचार्ड माइक्रोओरगनीजम. माइक्रोबायोलॉजि एंड मोलिकुलर बयोलॉजी रेवियूस, 68(4), 195-195.
- मेयर, एफ., परमान्न, डी., डी' सुजा, एम., ओलसन, आर., ग्लास, ई. एम., कुबल, एम., एंड विकेनिंग, जे. (2008). द मटाजेनोमिक्स रस्ट सर्वर-अ पब्लिक रिसौर्सेस फॉर द ऑटोमैटिक फ़ाईलोजिनैटिक एंड फंकशनल एनालिसिस ऑफ मटाजेनोमिक्स. बीएमसी बायोइन्फोर्मेटिक्स, 9(1), 386.
- लिन, एच. एच. एंड लियाओ, वाई. सी. (2016). एक्यूरेट बाईनिंग ऑफ मटाजेनोमिक्स कोण्टिज्स वाया औटोमटेड क्लास्ट्रिंग सीक्वेंसेस यूजिंग इन्फॉर्मेशन ऑफ जेनॉमिक सिग्नेचर एंड मार्कर जेनेस. साईंटिफ़िक रेपोर्ट्स, 6, 24175.

- पेंग, वाई., लेउंग, एच. सी. एम. विउ, एस. एम. चिन, एफ. वाई. एल. (2011). मेटा-आईडीबीए: अ डे नोवो असेंबलर फॉर मटाजेनोमिक्स डाटा. बायोइन्फोर्मेटिक्स, डोई: 10. 1093/बायोइन्फोर्मेटिक्स/बीटीआर216 पीएमआईडी: 21685107

- नामिकी, टी., हचिया, टी., तनका, एच., सकाकीबरा, वाई. (2012). मेटावेल्वेट: एन एक्सटैन्शन ऑफ वेल्वेट असेम्बलेर टु डे नोवो मटाजेनोमिक्स असेम्बलेर फ्रम शॉर्ट सीक्वेंस रिड्स. न्यूकलिएक एसिडस रेस., डोई: 10.1093/नर/जीकेएस678 पीएमआईडी: 22821567

- जेरबीनों, डी. आर. बिरने, ई. (2008). वेल्वेट: अलगोरीथम फॉर डे नोवो शॉर्ट रीड असेंबली यूजिंग डे बृइज्न ग्राफ्स. जेनोम रेस. डोई: 10. 1101/जीआर. 074492.107 पीएमआईडी: 1834938643.

- हैदर, बी., अहन, टी. एच., बशनेल, बी., चाय, जे., कोपेलेंड, ए., पैन, सी. (2014). ओमेगा: एन ओवेरलप-ग्राफ डे नोवो असेम्बलेर फॉर मटाजेनोमिक्स बायोइन्फोर्मेटिक्स; डोई: 10.1093/बायोइन्फोर्मेटिक्स/बीटीयू 395 पीएमआईडी: 25609793

- म्येर्स, ई. डब्लू. (2005). द फ्रगमेंट असेंबल स्ट्रिंग ग्राफ. बायोइन्फोर्मेटिक्स.; डोई: 10.1093/बायोइन्फोर्मेटिक्स/ बीटीयू 114 पीएमआईडी: 25609793

- ली, डी., लिउ, सी. एम., लुओ, आर., सदाकने, के. एंड लम, टी. डब्लू. (2015). मेगाहिट: अन अल्ट्रा-फास्ट सिंगल-नोड सोल्यूशंस फॉर लार्ज एंड कॉम्प्लेक्स असेंबल वाया सकसीनकट डे बृइज्न ग्राफ. बायोइन्फोर्मेटिक्स, डोई: 10. 1093/बायोइन्फोर्मेटिक्स/बीटीवी033 पीएमआईडी: 25609793

- हूसन एट अल (2011). इंटेग्रटिव एनालिसिस ऑफ एनवायरनमेंट सीक्वेंसेस यूजिंग एमईजीएएन4, जीनोम रेस., 21, 1552-1560.

- विलके, ए. बिछोफ, जे., गेरलच, डब्लू., ग्लास, ई., हैरिसन, टी., कीगन, के. पी., पाकजिआन, टी., त्रिंबल, डब्लू. एल., बगची, एस., ग्रामा, अ., एट अल. (2015). द एमजी-आरएएसटी मटाजेनोमिक्स डेटाबेस एंड पोर्टल इन 2015. न्यूकलेक एसिडस रेस., 44(डी1), 590-94.

- वूड, डी. ई. एंड साल्ज़बर्ग, एस. एल. (2014). करकें: अल्ट्राफास्ट मटाजेनोमिक्स सीक्वेंस क्लासिफिकेशन यूजिंग अगजेक्ट अलिग्न्मेंट, जेनोम बयोलॉजी, 15:आर46

- ब्रैंडी, ए., एंड साल्ज़बर्ग, एस. एल. (2009). फीमम एंड फीममबीएल: मटाजेनोम फीलोजेनेटिक क्लासिफिकेशन विद इंटेर्पोलेटिड मार्कोव मोडेसल. मेथड्स, 6(9), 673-676. हेप्ट://डोई.ऑर्ग/10.1038/नमेथ.1358

- मैकडोनाल्ड, एन. जे., पार्क्स, डी. एच., बेइको, आर. जी. (2012) रैपिड आईडेनटीफिकेशन ऑफ हाइ-कॉन्फ़िडेंस टेक्सोनोमिक असिग्नमेंट फॉर मटाजेनोमिक्स डाटा. न्यूकलेक एसिडस रेस., 40(14):ई111. डोई: 10.1093/एनएआर/जीकेएस335. एपब अपीआर 24.

- किसल्यूक, अ., भटनागर, एस., डूशोफ़्फ़, जे., एंड वेट्ज़, जे. एस. (2009). अनसुपरवाईजड स्टेटीससटिकल क्लस्टारींग ऑफ एनवायरनमेंट शॉटगन सीक्वेंसज. बीएमसी बायोइनफॉरमटिक्स, 10(1), 316.

कृषि में प्रोटिओमिक्स डेटा विश्लेषण का अवलोकन

भूमिका

प्रोटीन एक महत्वपूर्ण जैविक स्थूल अणु है जो कि विभिन्न प्रकार के कार्य करते हैं । शब्द "प्रोटिओम" को जीव द्वारा उत्पादित या संशोधित प्रोटीन के पूरे समूह के रूप में परिभाषित किया गया है । प्रोटिओमिक्स आमतौर पर किसी कोशिका प्ररूप के लिए प्रोटीन के बड़े पैमाने पर मात्रात्मक/ गुणात्मक अध्ययन को संदर्भित करता है । अब यह विभिन्न क्षेत्रों में एक शक्तिशाली टूल के रूप में उभरा है जैसे कि बायोमेडिसिन (मुख्य रूप से रोगों के लिए), कृषि और पशु विज्ञान । यह पौधों के कार्यों के विभिन्न पहलुओं के अध्ययन के लिए तेजी से महत्वपूर्ण होता जा रहा है जैसे कि कीटों से पौधों की रक्षात्मक प्रतिक्रिया में प्रत्याशी प्रोटीन की पहचान, फसल उत्पादन पर भू-मंडलीय मौसम परिवर्तन का प्रभाव, इत्यादि । प्रोटिओमिक्स के अनुप्रयोगों में प्रोटिओमिक्स एक्सप्रेसन/ एबंडेंस, संरचनात्मक प्रोटिओमिक्स, बायोमार्कर का आविष्कार, इंटरैक्शन प्रोटिओमिक्स, प्रोटीन नेटवर्क, आदि शामिल हैं ।

आमतौर पर प्रोटीन एक्सप्रेसन डेटा (आंकड़ा) मास स्पेक्ट्रोमीटर जैसे उच्च-थ्रूपुट तकनीक का उपयोग करके उत्पन्न किया जाता है । जटिल मिश्रण में प्रोटीन और पेप्टाइड्स की पहचान और परिमाणन के लिए प्रोटिओमिक्स में तरल क्रोमैटोग्राफी - मास स्पेक्ट्रोमेट्री (एम.एस.) का उपयोग एक विधि के रूप में किया जाता है । प्रोटिओमिक्स के दो बुनियादी दृष्टिकोण हैं, अर्थात् बॉटम-अप (नीचे-ऊपर) और टॉप-डाउन (ऊपर-नीचे) । सबसे आम प्रोटिओमिक्स दृष्टिकोण बॉटम-अप है जिसमें एक सैम्पल में प्रोटीन एंजाइमेटिक क्रिया से पेप्टाइड्स में टूट जाता है और उसके पश्चात क्रोमैटोग्राफिक पृथक्करण, आयनीकरण और मास विश्लेषण किया जाता है । इसके विपरीत, टॉप-डाउन प्रोटिओमिक्स पूर्ण प्रोटीन के अध्ययन को संबोधित करता है और अक्सर इसका उपयोग शुद्ध या आंशिक रूप से शुद्ध प्रोटीन के लिए किया जाता है । इसके अलावा, विभिन्न स्थितियों के बीच फीचर एबंडेंस में अंतर का पता लगाने के लिए फीचर्स (प्रोटीन अथवा पेप्टाइड्स) का परिमाणन लेबल-मुक्त या लेबल-युक्त (मेटाबोलिक, एंजाइमेटिक या रासायनिक) हो सकता है । लेबल-मुक्त परिमाणन में, फीचर्स की एम.एस. आयन तीव्रता और स्पेक्ट्रल गणना प्रमुख दृष्टिकोण हैं ।

प्रोटिओमिक्स आंकड़ों में मिसिंग वेल्युस एवं विषमता

प्रोटिओमिक्स डेटा विश्लेषण के लिए विभिन्न दृष्टिकोण मौजूद हैं, जिसमें पहला कदम फीचर्स की आयन तीव्रता को संक्षेप में प्रस्तुत करना है और इसके बाद कुछ परिवर्तन जैसे कि लॉग परिवर्तन का उपयोग कर उसे नॉर्मल डिस्ट्रिब्युसन के समीप लाते हैं । प्रोटिओमिक्स डेटा विश्लेषण के टूल्स/ विधियाँ की उपलब्धता के बावजूद, प्रोटिओमिक्स डेटा का विश्लेषण करने में विभिन्न सांख्यिकीय चुनौतियाँ हैं, जैसे कि डेटा विविधता और अनुपस्थित अवलोकन (मिसिंग वेल्यु) । प्रत्येक विधियों में कई कमियां हैं जिनका अध्ययन इन विधियों के सांख्यिकीय गुणों का अध्ययन करके किया जा सकता है ।

सैम्पल्स के बीच जैविक परिवर्तिता (वेरियेब्लीटी) और डेटा उत्पादन (जनरेशन) के तकनीकी दृष्टिकोण विषमता को जन्म देती है । जैविक परिवर्तिता आनुवंशिक और पर्यावरणीय कारकों से उत्पन्न होती है । तकनीकी दृष्टिकोण जैसे कि सैम्पल निकालना, संचयन, बफर के विभिन्न बैच, मास स्पेक्ट्रोमीटर रन को दोहराना, आदि एबंडेंस डेटा में परिवर्तन लाते हैं ।

जब एक डेटा सेट के प्रत्येक समूह में समान संख्या में सैम्पल (विषय) होते हैं, और जब फीचर्स में कोई अनुपस्थित अवलोकन (मिसिंग वेल्यु) नहीं होता है, तो उस डेटा सेट को संतुलित कहा जाता है । यह स्थिति हमेशा नहीं होती । डेटा असंतुलित भी हो सकता है, जिसमें सैम्पल की असमान संख्या, या मिसिंग वेल्यु या दोनों हो सकते हैं । प्रोटिओमिक्स डेटा में मिसिंग वेल्यु जैविक और / या तकनीकी मुद्दों के कारण हो सकते हैं । मिसिंग वेल्यु तीन प्रकार के होते हैं: (i) मिसिंग कम्पलिट्ली एट रैंडम (MCAR), जिसमें मिसिंग वेल्युस दोनों अप्रत्यक्ष और प्रत्यक्ष डेटा से स्वतंत्र होते हैं; (ii) मिसिंग एट रैंडम (MAR), जिसमें यदि प्रत्यक्ष डेटा पर सशर्त, मिसिंग वेल्युस मिसिंग माप से स्वतंत्र हैं; और (iii) मिसिंग नॉट एट रैंडम (MNAR), जब डेटा न तो MCAR है और न ही MAR है । मिसिंग वेल्युस वाले डेटा का विश्लेषण या तो मिसिंग वेल्युस वाली फीचर्स को हटा कर, या सांख्यिकीय तरीकों का उपयोग करके जो कि असंतुलित डेटा को हैंडल कर सकता है, या इंप्यूटेशन विधियों का उपयोग करके किया जा सकता है । यदि मिसिंग वेल्युस वाली फीचर्स को हटा दिया जाए, तो जानकारी की हानि होगी । इसलिए, मिसिंग वेल्युस को हैंडल करने वाले तरीकों का उपयोग, जैसे कि इंप्यूटेशन विधियां, आमतौर पर चुनी जाती है ।

प्रोटिओमिक्स एबंडेंस/ एक्सप्रेसन विश्लेषण

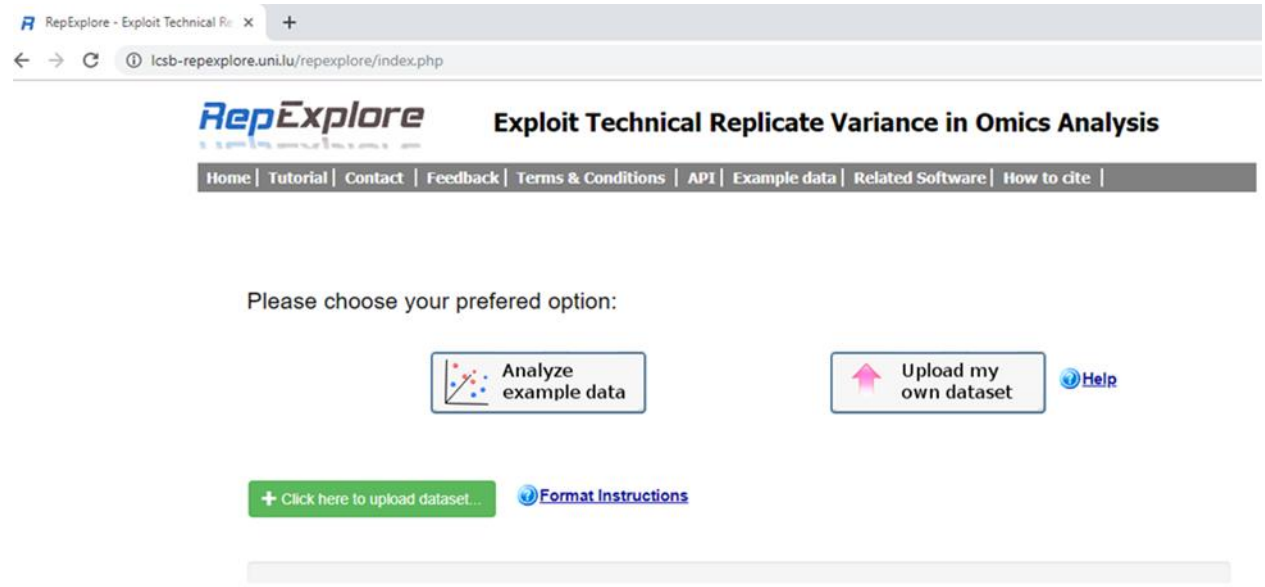
प्रोटिओमिक्स एबंडेंस विश्लेषण दो या दो से अधिक स्थितियों में महत्वपूर्ण फीचर्स का पता लगाने के लिए किया जाता है, जैसे कि स्वस्थ बनाम विभिन्न रोग स्थितियां । डेटा को विभिन्न परिवर्तन और/ या सामान्यीकरण विधियों [उदाहरण: लॉग परिवर्तन (logarithmic transformation), मात्रात्मक सामान्यीकरण (quantile normalization), विचरण स्थिर सामान्यीकरण (variance stabilizing normalization), आदि] का उपयोग करके सामान्यीकृत किया जाता है । यदि डेटा में मिसिंग वेल्यु है तो इंप्यूटेशन तकनीकों [उदाहरण: सिंगुलर वेल्यु डिकम्पोसिसन (singular value decomposition), *k*-निकटतम पड़ोसी (*k*-nearest neighbor), अधिकतम संभावना प्राक्कलन (maximum likelihood estimation), आदि] का उपयोग किया जा सकता है । महत्वपूर्ण फीचर्स की पहचान के लिए विभिन्न सांख्यिकीय पद्धतियां हैं जैसे टी-टेस्ट (t-test), मॉडरेटेड टी-टेस्ट (moderated t-test), अनोवा (ANOVA), लीनियर मिक्स्ड मॉडल (linear mixed model), इत्यादि ।

फीचर्स के अंतर एक्सप्रेसन विश्लेषण के लिए विभिन्न टूल्स और पैकेज उपलब्ध हैं जैसे कि “RepExplore”, “MSqRob”, “MSstats”, “PANDA”, इत्यादि । वेब सर्वर “RepExplore” का उपयोग करके प्रोटिओमिक्स एबंडेंस विश्लेषण का एक उदाहरण नीचे वर्णित है । केस-कंट्रोल अध्ययन के लिये एक टेस्ट डेटासेट का भाग उदाहरण के तौर पर नीचे दिया गया है (चित्र 1) । डेटासेट के दोनों समूह (केस और कंट्रोल) में 2 जैविक सैम्पल हैं तथा प्रत्येक जैविक सैम्पल के 2 तकनीकी रेप्लिकेट हैं ।

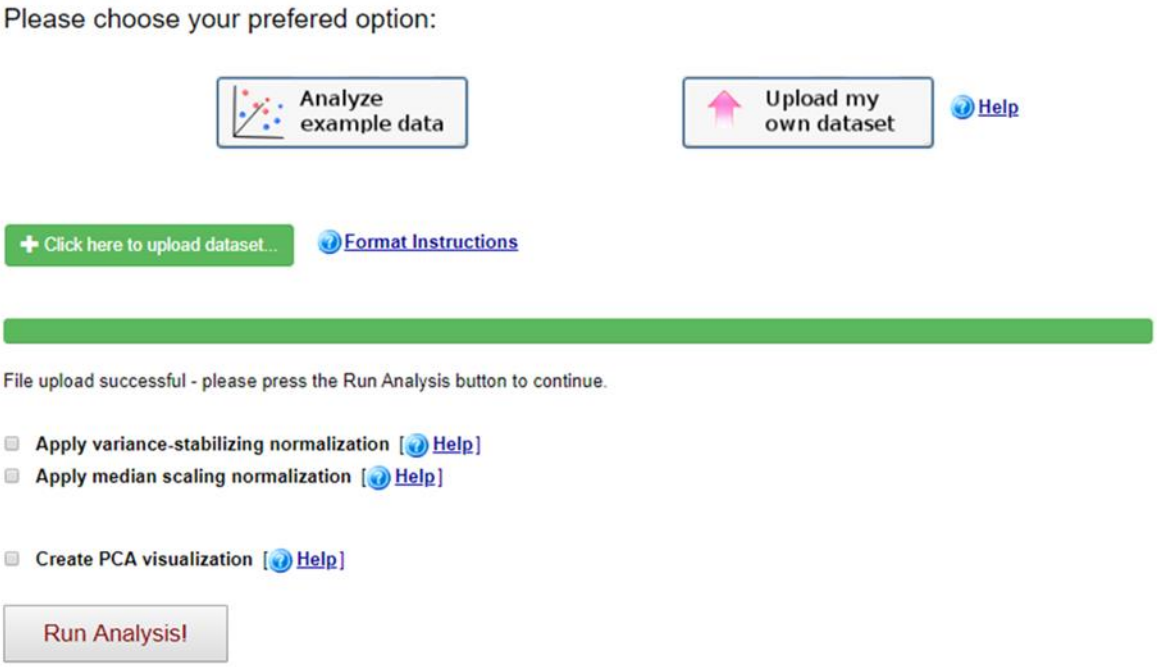
	Control				Case			
	control_1_1	control_1_2	control_2_1	control_2_2	case_1_1	case_1_2	case_2_1	case_2_2
biomolecule_1	20.84	19.93	20.78	19.24	20.03	20.87	19.65	20.07
biomolecule_2	19.18	18.79	18.88	18.43	18.97	18.88	18.82	18.64
biomolecule_3	19.5	18.84	20.14	19.06	19.58	19.29	19.1	19.31
biomolecule_4	19.23	18.52	19.67	17.73	19	18.6	16.4	18.44
biomolecule_5	19.64	19.25	19.99	18.78	19.5	19.31	19.16	19.41
biomolecule_6	19.89	19.45	19.93	18.8	19.46	18.76	18.84	18.94
biomolecule_7	22.07	21.72	23.26	21.35	22.74	21.65	20.97	22.17
biomolecule_8	21.84	21.47	22.81	21.22	22.35	21.58	21.18	22.01
biomolecule_9	17.56	17.41	17.46	17.7	16.76	18.13	18.51	17.3
biomolecule_10	20.34	19.81	21.02	19.23	20.38	19.6	19.06	19.8
biomolecule_11	19.15	18.79	17.98	19.03	17.81	19.55	19.89	18.76
biomolecule_12	24.64	24.12	23.21	24.38	23.31	24.77	25.04	24.21
biomolecule_13	26.51	26.06	26.74	25.23	26.32	25.67	25.15	25.95
biomolecule_14	25	24.42	23.27	24.79	23.48	25.16	25.45	24.58
biomolecule_15	18.05	18.3	18.51	17.98	18.52	17.4	18.36	16.85
biomolecule_16	17.82	17.34	18.24	17.07	17.8	17.66	17.45	17.65
biomolecule_17	17.98	17.31	18.28	17.27	17.77	17.37	17.47	17.31
biomolecule_18	19.32	18.13	19.33	18.04	18.81	18.64	18.66	17.88
biomolecule_19	24.89	24.43	24.82	23.74	24.4	24.24	24.1	24.35
biomolecule_20	17.94	17.25	18.39	17.19	17.19	17.08	16.88	16.82

चित्र 1. केस-कंट्रोल अध्ययन के लिये टेस्ट डेटासेट का भाग

उपयोगकर्ता को डेटा अपलोड करना होता है (चित्र 2) । डेटा अपलोड करने के बाद, उपयोगकर्ता को अन्य विकल्पों का चयन करना होता है (चित्र 3) । इसके बाद उपयोगकर्ता को “Run Analysis!” बटन पर क्लिक करना है (चित्र 3) ।

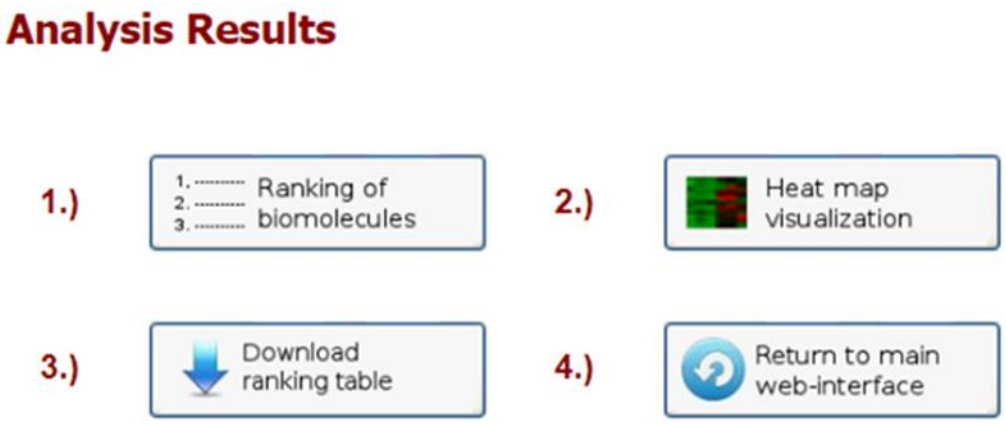


चित्र 2. डेटा अपलोड करना



चित्र 3. विकल्पों का चयन

तब उपयोगकर्ता को विभिन्न परिणाम मिलते हैं जो कि नीचे दिए गए हैं (चित्र 4-6) ।



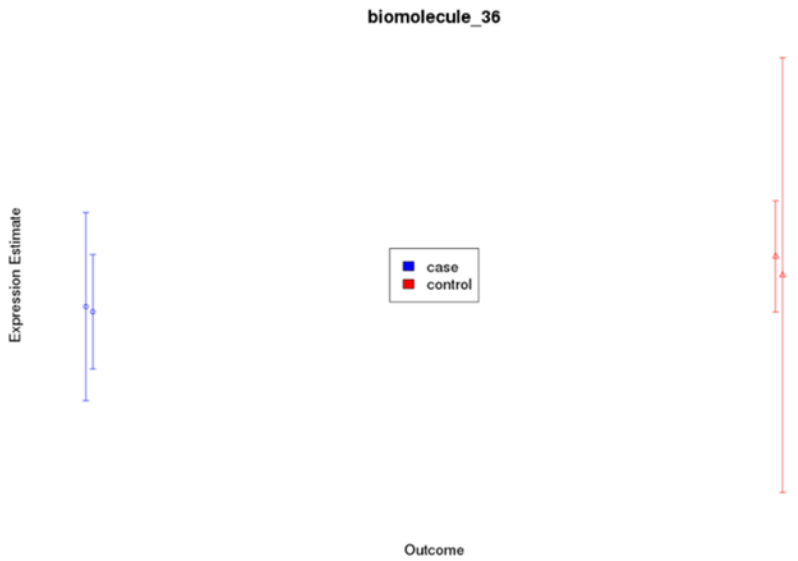
चित्र 4. विश्लेषण के विभिन्न परिणाम के मेन्यू

डिफ्रेंसियली एबंडेंट/ एक्सप्रेस्ड फीचर्स की रैंकिंग टेबल चित्र 5 में दी गई है ।

Biomolecule identifier	Log. fold change	Probability of positive likelihood ratio (PPLR)	P-like significance score (min(PPLR,1-PPLR))	eBayes T-score	eBayes P-value	eBayes adj. P-value
biomolecule_90 generate bar plot	1.66	0.287	0.287	3.71	0.0116	0.542
biomolecule_36 generate bar plot	-0.8	0.633	0.367	-3.27	0.0193	0.542
biomolecule_20 generate bar plot	-0.7	0.762	0.238	-2.88	0.0309	0.542
biomolecule_93 generate bar plot	0.858	0.311	0.311	2.85	0.032	0.542
biomolecule_100 generate bar plot	0.93	0.378	0.378	2.85	0.0322	0.542
biomolecule_40 generate bar plot	0.843	0.236	0.236	2.79	0.0345	0.542
biomolecule_94 generate bar plot	-0.943	0.672	0.328	-2.72	0.0379	0.542
biomolecule_73 generate bar plot	-0.517	0.69	0.31	-2.22	0.0718	0.586

चित्र 5. डिफ्रेंसियली एबंडेंट/ एक्सप्रेस्ड फीचर्स की रैंकिंग

उपयोगकर्ता किसी भी फीचर का बार प्लॉट "generate_bar_plot" बटन पर क्लिक करके प्राप्त कर सकता है जिसका एक उदाहरण नीचे दिया गया है ।



चित्र 6. एक फीचर के बार प्लॉट का उदाहरण

निष्कर्ष

यह लेख प्रोटिओमिक्स और जैव सूचना विज्ञान के क्षेत्र में काम करने वाले शोधकर्ताओं के लिए उपयोगी है । हमने प्रोटिओमिक्स एबंडेंस डेटा का विश्लेषण करने के लिए अलग-अलग तरीकों का उल्लेख किया है । इसके अलावा, इन तरीकों का उपयोग समान प्रयोगों के एक्सप्रेसन डेटा (जैसे, माइक्रोएरे और मेटाबॉलोमिक्स डेटा) का विश्लेषण करने के लिए किया जा सकता है ।

संदर्भ

एंडरसन एन.एल. और एंडरसन एन.जी. (1998). प्रोटीन और प्रोटिओमिक्स: नई तकनीक, नई अवधारणाएँ और नए शब्द । *इलेक्ट्रोफोरेसिस,* **19 (11),** 1853-1861 ।

करपीविच वाई.वी., डबनी ए.आर. और स्मिथ आर.डी. (2012). लेबल-मुक्त LC-MS विश्लेषण के लिए सामान्यीकरण और मिसिंग वेल्यु इंप्यूटेशन । *बी.एम.सी. बायोइन्फोर्मेटिक्स,* **13 सप्ल 16**, S5 ।

रुबिन डी.बी. (1976). इंफेरेंस और मिसिंग डेटा । बायोमेट्रिका, **63 (3)**, 581-592 ।

ग्लैब ई. और श्नाइडर आर. (2015). RepExplore: प्रोटिओमिक्स और मेटाबॉलिकम डेटा विश्लेषण में तकनीकी रेप्लिकेट वैरियेंस को संबोधित करते हुए । *बायोइन्फोर्मेटिक्स,* **31 (13),** 2235-7 ।

गोएमीन एल.जे.ई., गेवर्ट के., और क्लेमेंट एल. (2018). लेबल-मुक्त मात्रात्मक LC / MS प्रोटिओमिक्स में प्रयोगात्मक डिज़ाइन और डेटा-विश्लेषण: MSqRob के साथ एक ट्यूटोरियल । *ज. प्रोटिओमिक्स,* **171**, 23-36 ।

चांग सी., एट अल. (2018). PANDA-view: सांख्यिकीय विश्लेषण और मात्रात्मक प्रोटिओमिक्स डेटा के विजुआलाइज़ेसन के लिए एक उपयोग करने में आसान वाला टूल । *बायोइन्फोर्मेटिक्स* ।

चोई एम., एट अल (2014). MSstats: मात्रात्मक द्रव्यमान स्पेक्ट्रोमेट्री आधारित प्रोटिओमिक प्रयोगों के सांख्यिकीय विश्लेषण के लिए एक आर पैकेज । *बायोइन्फोर्मेटिक्स,* **30 (17),** 2524-6 ।

प्रोटीन संरचना मॉडलिंग

भूमिका

कम्प्यूटेशनल बायोलॉजी के क्षेत्र में स्ट्रक्चरल बायोइनफॉरमैटिक्स प्रमुख अनुसंधान क्षेत्रों में से एक है। स्ट्रक्चरल बायोइनफॉरमैटिक्स प्रोटीन, आर.एन.ए. और डी.एन.ए. जैसे जैविक अणुओं की 3-डी (त्रिआयामी) संरचनाओं के विश्लेषण और पूर्वानुमान/ भविष्यवाणी को दर्शाता है । इस स्ट्रक्चरल जानकारी में प्रोटीन क्रिस्टलोग्राफी, इलेक्ट्रॉन माइक्रोस्कोपी या एनएमआर जैसे विभिन्न प्रयोगात्मक विधियों के माध्यम से प्राप्त 3-डी मैक्रोमोलेक्युलर संरचनाओं से मेल खाती है। इस जानकारी को प्रोटीन, आणविक तह, और विकास और संरचना के अध्ययन में उपयोग कर सकते है । स्ट्रक्चरल बायोइनफॉरमैटिक्स में त्रिआयामी प्रोटीन संरचनाओं का पूर्वानुमान, मुख्य शोध समस्याओं में से एक है।

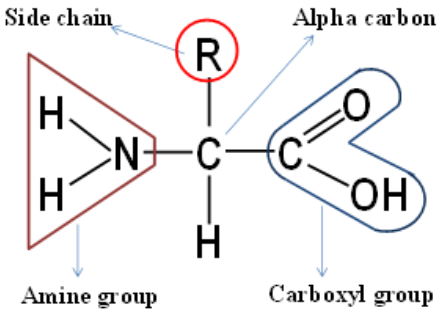
अधिक लागत एवम् समय उपभोग की दृष्टी से प्रोटीन संरचना का निर्धारण प्रयोगात्मक विधियों (क्रिस्टलोग्राफी, इलेक्ट्रॉन माइक्रोस्कोपी या एन.एम.आर.) से उपयुक्त नहीं है (गुनेटरट, 2004)। प्रोटीन की 3-डी संरचना को निर्धारित करने और खोजने में कठिनाई ने जेनोम प्रोजेक्ट्स द्वारा उत्पन्न डेटा और प्रोटीन की 3-डी संरचनाओं जो वर्तमान में ज्ञात हैं की संख्या के बीच एक बड़ी विसंगति उत्पन्न की है, । प्रयोगात्मक रूप से प्रोटीन अनुक्रमों के केवल एक छोटे हिस्से की 3-डी संरचना ज्ञात है। ये आंकड़े केवल इसकी आवश्यकता की व्याख्या नहीं करते , बल्कि कम्प्यूटेशनल प्रोटीन संरचना पूर्वानुमान विधियों में भी आगे के शोध को प्रेरित करते हैं। पिछले 10 वर्षों में कई कम्प्यूटेशनल पद्धतियों, सिस्टम और एल्गोरिदम को 3-डी प्रोटीन संरचना पूर्वानुमान (3-डी पीएसपी) समस्या (ओस्गॉप्पर, 2000) के समाधान के रूप में प्रस्तावित किया गया है।

यदि प्रोटीन की माध्यमिक (माध्यमिक), तृतीयक और चतुर्भागात्मक संरचना की भविष्यवाणी उसके अमीनो एसिड अनुक्रम (सीक्वेंस) अर्थात प्राथमिक संरचना से की जाए तो उसे प्रोटीन संरचना भविष्यवाणी कहते हैं। यह संरचनात्मक जीव विज्ञान (स्ट्रक्चरल बायोलॉजी) में सबसे महत्वपूर्ण लक्ष्यों में से एक है। प्रोटीन के कार्य की बेहतर समझ के लिए प्रोटीन की 3-डी संरचना को जानना महत्वपूर्ण है। प्रोटीन संरचना भविष्यवाणी के लिए कई दृष्टिकोण हैं और द्विवार्षिक CASP प्रयोग (प्रोटीन संरचना भविष्यवाणी के लिए तकनीकों का महत्वपूर्ण मूल्यांकन) में वर्तमान तरीकों के प्रदर्शन का आकलन किया जाता है।

प्रोटीन

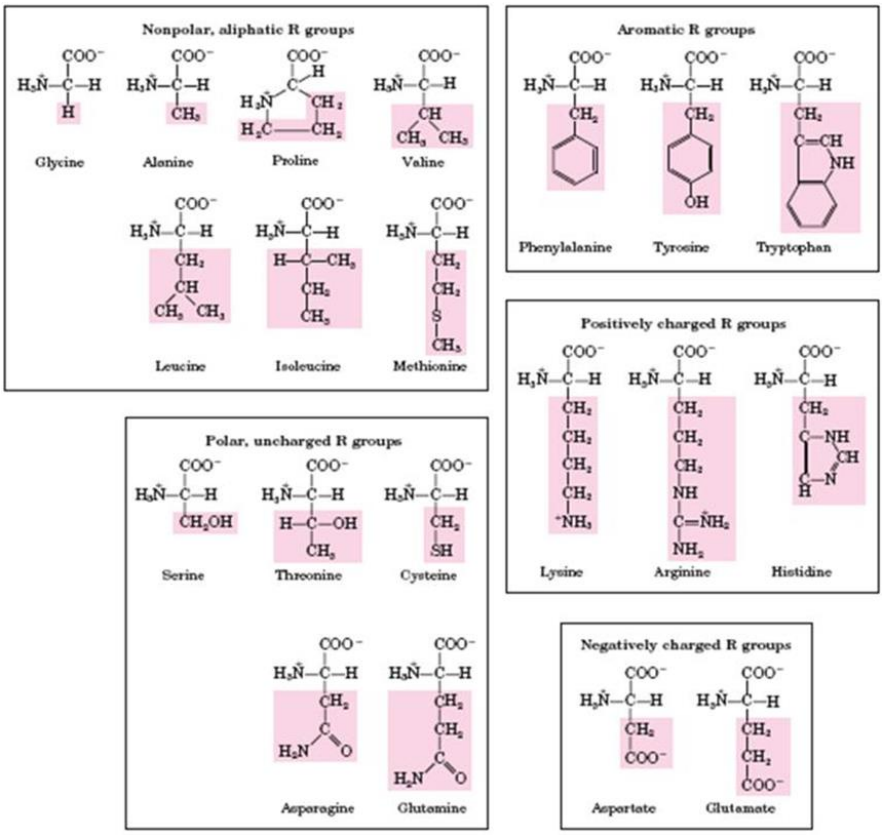
प्रोटीन नियमित, लीनियर पॉलिमर हैं जो अमीनो एसिड से बने होते हैं। प्रोटीन जीवन के लिए आधारभूत हैं । पॉलीसेकेराइड और न्यूक्लिक एसिड जैसे अन्य जैविक मैक्रोमोलेक्यूल्स की तरह, प्रोटीन कोशिकाओं के भीतर लगभग हर प्रक्रिया में भाग लेते हैं। उनके पास उल्लेखनीय कार्य हैं। उदाहरण के लिए, वे अधिकांश जैव रासायनिक प्रतिक्रियाओं को उत्प्रेरित करने वाले एंजाइम के रूप में कार्य करते हैं। प्रोटीन में संरचनात्मक या यांत्रिक कार्य भी होते हैं, जैसे मांसपेशियों में एक्टिन और मायोसिन और साइटोस्केलेटन में प्रोटीन, जो मचान की एक प्रणाली बनाते हैं जो कोशिका के आकार को बनाए रखता है। अन्य प्रोटीन कोशिका संकेतन, प्रतिरक्षा प्रतिक्रिया, कोशिका आसंजन और कोशिका चक्र में महत्वपूर्ण हैं। जानवरों के आहार में भी प्रोटीन आवश्यक है, क्योंकि जानवर उन सभी अमीनो एसिड को संश्लेषित (सिंथेसाइज़) नहीं कर सकते हैं जिनकी उन्हें आवश्यकता होती है। जानवरों में पाचन की प्रक्रिया के माध्यम से, प्रोटीन को फ्री अमीनो एसिड में तोड़ दिया जाता है तथा उनका उपयोग चयापचय में किया जाता है। संक्षेप में, लगभग हर जैविक प्रक्रिया में प्रोटीन का केंद्रीय महत्व होता है।

एमिनो एसिड: एक एमिनो एसिड, जिसे प्रोटीन में रेसिड्यु भी कहा जाता है, एक कार्बोक्सिल समूह, एक एमिनो समूह और एक साइड चेन से बना है (चित्र 1)।

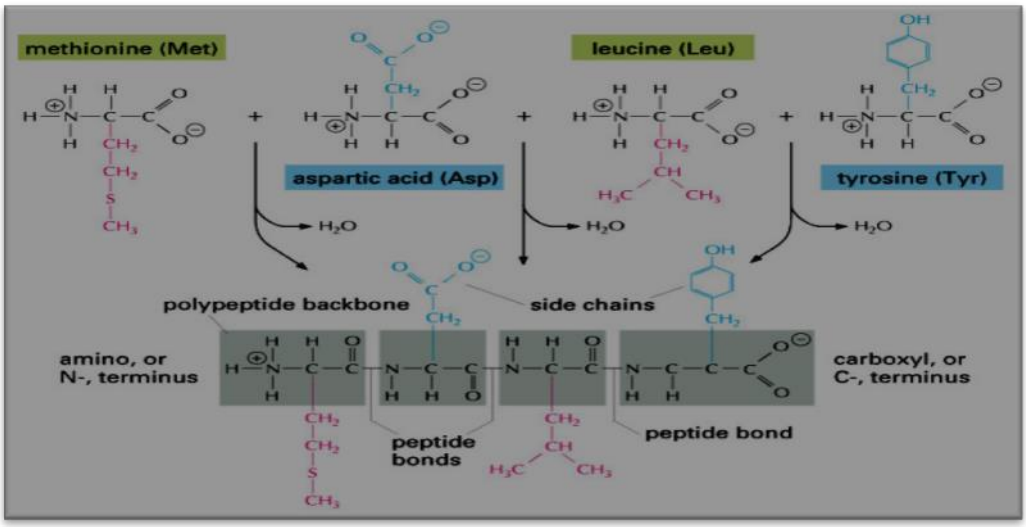


चित्र 1. अमीनो एसिड की सामान्य संरचना

प्रकृति में 20 अमीनो एसिड होते हैं जो हर प्राकृतिक प्रोटीन के लिए आधार बनाते हैं और ये केवल साइड चेन परमाणुओं में भिन्न होते हैं जैसा कि चित्र 2 में दिखाया गया है (लेहनिंगर *एट अल*, 2005)। अमीनो एसिड दो सटे हुए रेसिड्यु के अमीनो और कार्बोक्सिल समूहों के बीच पेप्टाइड बॉन्ड के गठन से प्रोटीन में जुड़े होते हैं (कृपया चित्र 3 देखें)।



चित्र 2. प्रोटीन के 20 आम अमीनो एसिड



चित्र 3. पेप्टाइड बांड गठन

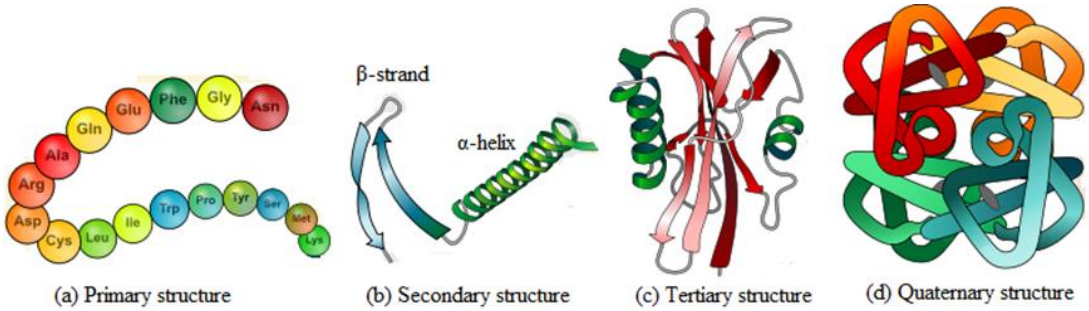
प्रोटीन संरचना के स्तर (लेवल्स): प्रोटीन संरचनाओं को निम्नलिखित तरीकों से परिभाषित किया गया है (चित्र 4) (लेहनिंगर *एट अल*, 2005)।

(a) प्राथमिक संरचना: यह पॉलीपेप्टाइड श्रृंखला बनाने वाला एमिनो एसिड सीक्वेंस है।

(b) द्वितीयक (माध्यमिक) संरचना: यह स्थानीय संरचनात्मक पैटर्न, α-हेलिक्स, β-स्ट्रैंड (β-शीट में इकट्ठा हुए स्ट्रैंड्स के समूह), टर्न और इंटरकनेक्टिंग लूप द्वारा परिभाषित किया गया है।

(c) तृतीयक संरचना: प्रोटीन की 3-डी संरचना तृतीयक संरचना है। यह प्रोटीन का कार्यात्मक रूप है।

(d) चतुर्धातुक संरचना: प्रोटीन का एकत्रीकरण और जटिल गठन दो या अधिक पॉलीपेप्टाइड सबयूनिट्स चतुर्धातुक संरचना को परिभाषित करता है।



चित्र 4. प्रोटीन में संरचनाओं का स्तर

3-डी प्रोटीन संरचना की भविष्यवाणी के तरीके

केवल एमिनो एसिड सीक्वेंस पर आधारित प्रोटीन की 3-डी संरचना की भविष्यवाणी आखिरी दशकों में, जैव रसायन, जीवविज्ञानियों, कंप्यूटर वैज्ञानिकों और गणितज्ञों के लिये एक चुनौती रही है। प्रोटीन संरचना पूर्वानुमान स्ट्रक्चरल बायोइनफॉरमैटिक्स में मुख्य शोध समस्याओं में से एक है (क्रेइटन, 1990)। मुख्य चुनौती यह है कि कैसे अमीनो एसिड अवशेषों के रेखीय अनुक्रम में एन्कोड की गई जानकारी को 3-डी संरचना में अनुवादित किया गया जाएँ, और इस अधिग्रहीत ज्ञान से ऐसि कम्प्यूटेशनल पद्धतियां विकसित की जाएँ जो सही तरीके से प्रोटीन के मूल संरचना का अनुमान लगा सके। इस जटिल समस्या का हल के रूप में कई तरीकों और एल्गोरिदम का प्रस्ताव, परीक्षण और विश्लेषण किया गया है। साहित्य में, कई 3-डी प्रोटीन संरचना भविष्यवाणी विधियों के कई वर्गीकरण उपलब्ध है। यहा पर हम, फ्लोडास (फ्लोडास *एट अल*, 2006) द्वारा वर्णित कम्प्यूटेशनल वर्गीकरण का अध्ययन करेंगे, जो कि प्रोटीन संरचना की भविष्यवाणी को चार समूहों में वर्गीकृत करता है:

1. डेटाबेस सूचना के बिना प्रथम सिद्धांत विधि

2. डेटाबेस सूचना के साथ प्रथम सिद्धांत विधि

3. फोल्ड पहचान और थ्रेडिंग विधियां

4. तुलनात्मक मॉडलिंग विधियां और अनुक्रम संरेखण रणनीति

डेटाबेस सूचना के बिना प्रथम सिद्धांत विधि

एब इनिसियो तरीके, डाटाबेस सूचना के बिना प्रथम सिद्धांत, थर्मोडायनामिक्स पर आधारित हैं जोकि इस तथ्य पर आधारित हैं कि प्रोटीन का मूल संरचना की ऊर्जा न्यूनतम हो (ट्रामोंटोनो, 2006)। *एब इनिसियो* संरचना भविष्यवाणी के तरीकों का लक्ष्य केवल एक एमिनो एसिड अनुक्रम पर आधारित, प्रोटीन की मूल संरचना की भविष्यवाणी करना है। शुद्ध एब इनिसियो विधियों में एक डाटाबेस से संरचनात्मक टेम्प्लेट्स का उपयोग जैसे कि पीडीबी की अनुमति नहीं है। एब इनिसियो प्रोटीन फोल्डिग को एक सार्वत्रिक अनुकूलन समस्या माना जाता है, जहां लक्ष्य को एक चर सेट के मूल्यों की पहचान करना है (टॉरसोन एंगल्स, सभी परमाणुओं की स्थिति या प्रोटीन संरचना में परमाणुओं का एक विशिष्ट सेट) जो कि पॉलीपेप्टाइड रचना की न्यूनतम ऊर्जा का वर्णन करते हैं।

एब इनिसियो पद्धतियां एक ऊर्जा समारोह का प्रयोग करके प्रोटीन रचना स्पेस का अनुकरण करती है, जो कि प्रोटीन की आंतरिक ऊर्जा का वर्णन करती है और जिस वातावरण में इसे डाला जाता है। इसका लक्ष्य सार्वत्रिक स्तर की निशुल्क ऊर्जा का पता लगाना है जो प्रोटीन के मूल या कार्यात्मक अवस्था (ट्रामोंटानो, 2006) से मेल खाती है। एबी इनिसियोपद्धतियां नए फोल्ड की भविष्यवाणी कर सकती हैं क्योंकि वे पीडीबी से टेम्पलेट तक सीमित नहीं हैं। हालांकि, कनफर्मेसनल सर्च स्पेस के संबंध में इन विधियों में कुछ सीमाएं हैं।

डेटाबेस सूचना के साथ प्रथम सिद्धांत विधि

डाटाबेस सूचना के साथ प्रथम सिद्धांत की विधि में प्रोटीन संरचनाओं के सामान्य नियम प्रोटीन डाटाबेस से लिये जाते हैं जोकि 3-डी प्रोटीन संरचनाओं के निर्माण में मदद करते हैं। ये पद्धति किसी ज्ञात संरचना को टारगेट अनुक्रम से तुलना नहीं करते हैं, बल्कि उनके छोटे छोटे टुकड़ों से करते हैं, अर्थात् ज्ञात प्रोटीन संरचनाओं (फ्लोडास एट अल, 2006) की तुलना टारगेट अनुक्रम के टुकड़ों से किया जाता हैं। यह अवलोकन से सामने आता है कि जब एक नया फोल्ड खोजां जाता है, तो यह ज्ञात संरचनाओं (ट्रामोंटानो, 2006) वाले प्रोटीन के संरचनात्मक मोटिफ से बनता है। इस प्रकार, यदि प्रोटीन फ्रेग्मेंट समान संरचनाओं में फोल्ड होते हैं, तो इस जानकारी या इन टुकड़ों का उपयोग प्रोटीन के 3-डी संरचनात्मक मॉडल बनाने के लिए किया जा सकता है। यह टुकड़ों पर आधारित तरीकों का सार है। एक प्रोटीन की रचना विभिन्न संरचनात्मक रूपों का प्रतिनिधित्व करने वाले अमीनो एसिड अनुक्रम के विभिन्न टुकड़ों के एक समूह के रूप में देखी जाती है जो एक 3-डी प्रोटीन संरचना बनाने के लिए संयुक्त हैं। यहां होमोलोग फ्रेग्मेंट की पहचान कि जाती हैं तथा उन्हें स्कोरिंग फँगसन और अनुकूलन एल्गोरिदम के माध्यम से संकलित करते है। इन फ्रेगमेंटस को फ्रेगमेंट संकलन के माध्यम से संकलित किया जाता है, जिसकी संरचना सबसे कम संभावित ऊर्जा के साथ मिलती है (सिमंस *एट अल*, 1997)। जब हमारा उद्देश्य सबसे कम ऊर्जा वाले पॉलीपेप्टाइड संरचना को ढूंढना हैं , तो ये विधियां एबी इनिसियो विधियों के समान होती हैं। हालांकि, उन्हें एब इनिसियो विधियों के रूप में वर्गीकृत नहीं किया जा सकता

क्योंकि वे पॉलीपीप्टाइड की संरचना का अनुमान लगाने के लिए डेटाबेस जानकारी का उपयोग करते हैं। आम तौर पर, जब प्रोटीन में अमीनो एसिड का पूरा अनुक्रम ज्ञात हो, एक फ्रेग्मेंट-आधारित विधि पांच विशिष्ट चरणों से मिलकर बनी है:

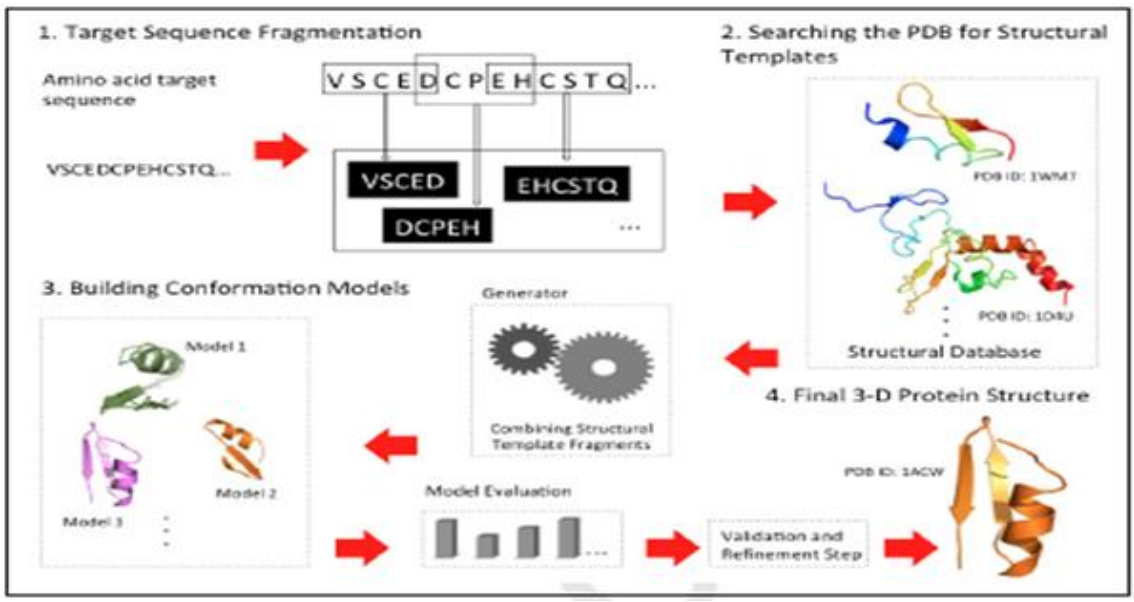
1. यह लक्षित अनुक्रम को फ्रेगमेंटस में विभाजित करता है

2. यह ज्ञात संरचना डेटाबेस में, प्रत्येक फ्रेगमेंट के समान अनुक्रमों की तलाश करता है

3. यह फ्रेगमेंटस को वर्गीकृत करता है (स्कोरिंग)

4. यह एक संयोजन तकनीकी का उपयोग करके टेम्पलेट फ्रेगमेंट से 3-डी संरचना का निर्माण करता है

5. संरचना परिष्करण (Refinement)

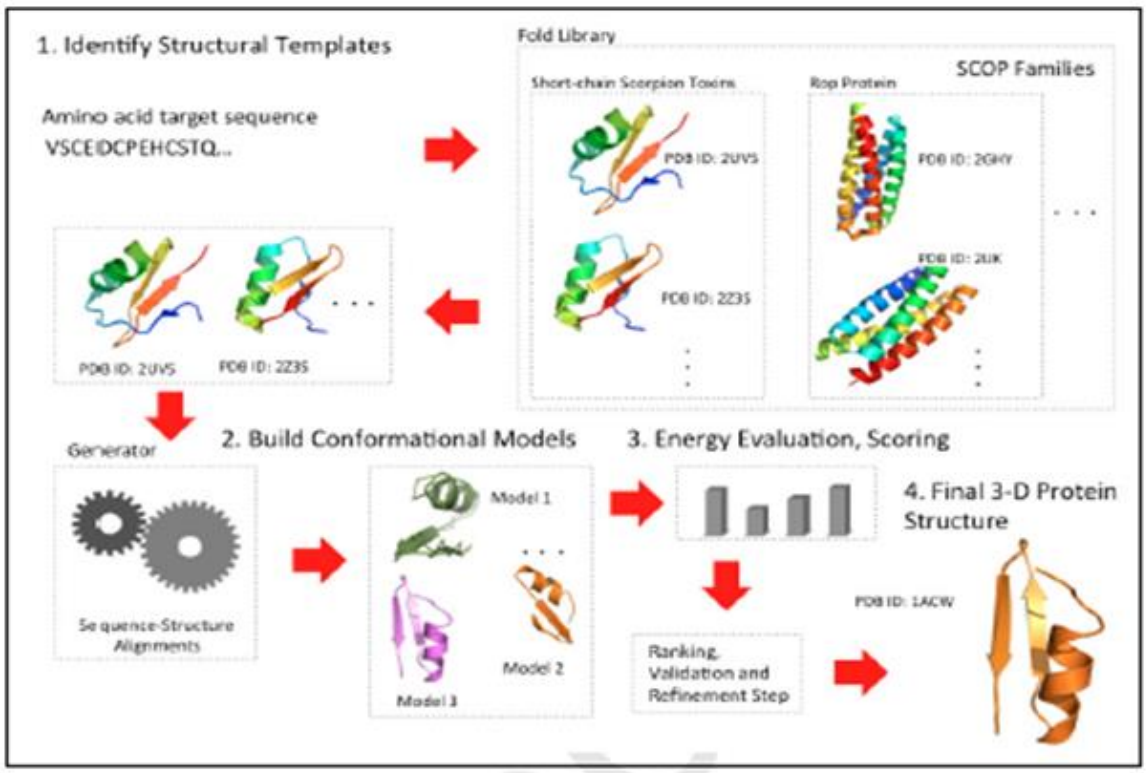


चित्र 5. 3-डी पीएसपी समस्या के लिए एक फ्रेगमेंट आधारित पद्धति का सामान्य योजनाबद्ध प्रतिनिधित्व

फोल्ड पहचान एवम् थ्रेडिंग आधारित विधियां

यह इस धारणा से प्रेरित हैं कि संरचना, अनुक्रम से अधिक संरक्षित है, अर्थात्, कोई स्पष्ट अनुक्रम समानता वाले प्रोटीन भी समान फोल्ड के हो सकते हैं (लेविट और चोथिया, 1976; फ्लौडास एट *अल*, 2006)। पिछले वर्षों में कई अध्ययनों से संकेत मिलता है कि प्रकृति में संरचनात्मक फोल्ड की संख्या सीमित है। उदाहरण के लिए, आज ज्ञात संरचना (रसेल और बार्टन, 1994) के प्रोटीन के पचास प्रतिशत भाग में लगभग दस अलग-अलग फोल्ड हैं। 3-डी प्रोटीन संरचना की थ्रेडिंग विधियों द्वारा पूर्वानुमान का लक्ष्य एक प्रोटीन अनुक्रम को किसी ज्ञात संरचनात्मक मॉडल पर फिट करना है। इस प्रक्रिया के दौरान लक्षित एमिनो एसिड अनुक्रम को उनके अनुक्रमिक क्रम के अनुसार टेम्पलेट 3 डी

संरचना पर अनुकूलतम तरीके से रखा जाता है। इसमें दो मूल प्रक्रियाएं शामिल हैं: (a) मॉडल की लाइब्रेरी से संरचनात्मक मॉडल का चयन करना और (b) संभावित अनुक्रम संरेखण के द्वारा संरचनात्मक मॉडल के प्रतिकूल लक्ष्य अनुक्रमों के बीच सही प्रतिस्थापन का पता लगाना। थ्रेडिंग विधियों संरचनात्मक जानकारी जैसे कि रेसिड्यु-रेसिड्यु संपर्क पैटर्न, माध्यमिक संरचना और विलायक पहुंचता, और संरचनात्मक समानताओं की पहचान करने के बाद, जोकि पूरी तरह से एमिनो एसिड अनुक्रमों के बीच समानता से नहीं पहचाना जा सकता है, इस तरह पूर्वानुमानित संरचनात्मक मॉडल का निर्माण किया जाता है।



चित्र 6. थ्रेडिंग प्रक्रिया का सामान्य योजनाबद्ध प्रतिनिधित्व

तुलनात्मक मॉडलिंग विधियां तथा अनुक्रम संरेखण रणनीतियाँ

तुलनात्मक मॉडलिंग विधि में अमीनो एसिड के लक्षित अनुक्रम को अन्य प्रोटीन के जिसकी अनुक्रम की संरचना ज्ञात हो जोकि और पीडीबी में संग्रहीत होता है, के साथ संरेखण करते (बर्मन *एट अल*, 2000)। यदि लक्षित अनुक्रम टेम्प्लेट प्रोटीन के अनुक्रम के समान है, तो ज्ञात संरचना से प्राप्त संरचनात्मक जानकारी का उपयोग लक्षित प्रोटीन के मॉडलिंग के लिए किया जाता है। इस प्रकार की विधि का मुख्य विचार है कि इसके एमीनो एसिड अनुक्रम से लक्षित प्रोटीन का एक परमाणु-रिज़ॉल्यूशन मॉडल तैयार करना। तुलनात्मक मॉडलिंग तभी लागू किया जा सकता है जब लक्षित प्रोटीन और टेम्पलेट प्रोटीन के बीच एक विकासवादी संबंध हो तथा जिसकी 3-डी संरचना ज्ञात हो। प्रोटीन के बीच विकासवादी संबंध तुलनात्मक मॉडलिंग विधियों का एक बुनियादी कारक है।

दिए गए एक प्रोटीन अनुक्रम के लिए, तुलनात्मक मॉडलिंग प्रक्रिया को होमोलोगस अनुक्रम ज्ञात संरचना के साथ, टेम्पलेट अनुक्रमों के प्रतिकूल query का संरेखण, अंतिम चरण में 3-डी मॉडल का निर्माण और शोधन की आवश्यकता होती है। इसके चार बुनियादी कदम इस प्रकार है:

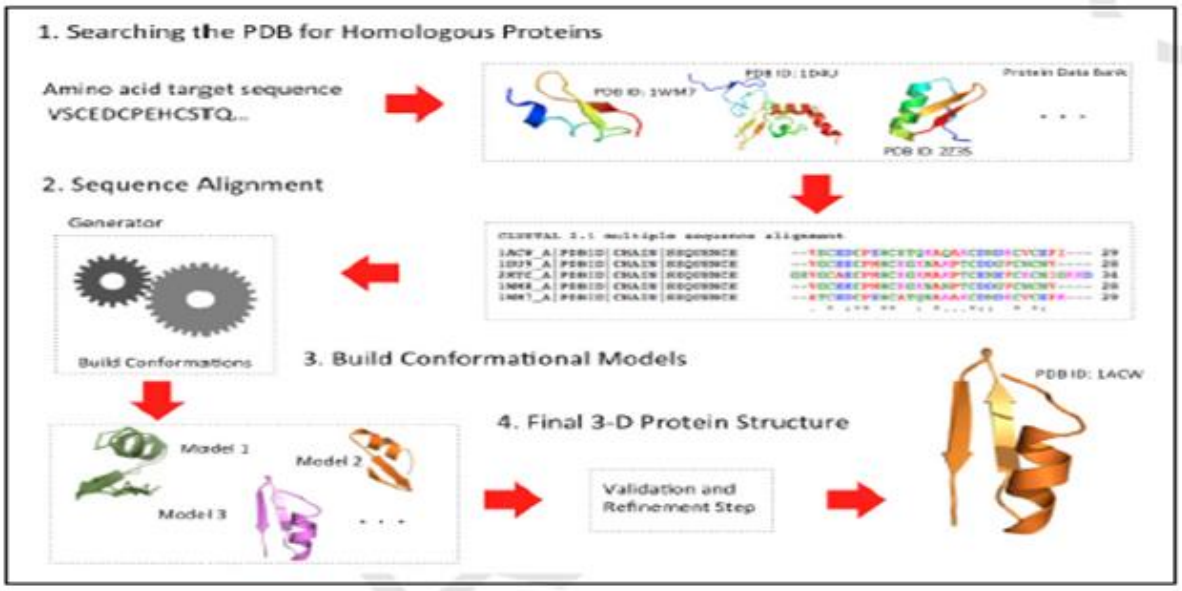
तुलनात्मक मॉडलिंग प्रक्रिया:

(1) फोल्ड असाइनमेंट तथा टेम्पलेट चयन

(2) टेम्पलेट लक्ष्य संरेखण

(3) मॉडल निर्माण

(4) मॉडल मूल्यांकन और परिशोधन



चित्र 7. तुलनात्मक मॉडलिंग की प्रक्रिया का योजनाबद्ध चित्रण

निष्कर्ष

स्ट्रक्चरल बायोइनफॉरमैटिक्स में प्रोटीन संरचना का अध्ययन और उनके 3-डी संरचनाओं का पूर्वानुमान महत्वपूर्ण अनुसंधान समस्याओं में से एक है। प्रोटीन संरचना का पूर्वानुमान और भी कठिन हो जाता हैं, यदि डेटा बैंक में कोई सम्बन्धित टेम्पलेट नहीं है। पिछले वर्षों में, इस जटिल समस्या को सुलझाने के उद्देश्य से कई कम्प्यूटेशनल विधियां, सिस्टम और एल्गोरिदम विकसित किए गए हैं। हालांकि, यह समस्या अभी भी जीवविज्ञानी, रसायनज्ञों, कंप्यूटर वैज्ञानिकों और गणितज्ञों केलिये चुनौती है। प्रोटीन संरचना का पूर्वानुमान एक बहुत ही मुश्किल समस्या है और आगे भी इसमें शोध की बहुत संभावना है। नई रणनीतियों का विकास, नए तरीकों की अनुकूलन और जांच और मौजूदा

अत्याधुनिक कम्प्यूटेशनल विधियों और 3-डी पीएसपी समस्या के संयोजन की स्पष्ट रूप से जरूरत है। एब इनिसियो तकनीक के साथ प्रयोगात्मक डेटा का बेहतर इस्तेमाल कैसे किया जा सकता है यह समझना एक मुख्य अनुसंधान का विषय है। संक्षेप में, कंप्यूटर विज्ञान, जैव सूचना विज्ञान, रसायन विज्ञान, जैव रसायन, और चिकित्सा विज्ञान जैसे बहुआयामी विभागों के अनुप्रयोग से इस क्षेत्र में कई शोध के अवसरों को खोजा जा सकता है।

संदर्भ

- बर्मन एच., वेस्टब्रुक जे., फेंग जेड., गिलिलैंड जी., बाथ टी., वीसिग एच., शिंद्यालोव आई., बॉर्न पी. (2000). प्रोटीन डाटा बैंक. *न्यूक्लिक एसिड रिसर्च*, 28 (1), 235।

- बक्सिनेसिस ए., क्विलैलेट बी. (1990). बायोइनफॉरमैटिक्स: ए प्रैक्टिकल गाइड टू दॅ एनालिसिस ऑफ जीन्स एंड प्रोटीन, द्वितीय संस्करण जॉन विले एंड सन्स, न्यूयॉर्क।

- फ्लोडास सी., फंग एच., मैकलिस्टर एस., मोनेगीमान एम., राजगरिया आर. (2006). प्रोटीन संरचना की भविष्यवाणी में प्रगति और प्रोटीन डिजाइन की नई जानकारी: एक समीक्षा. *रसायन अभियांत्रिकी विज्ञान*, 61 (3), 966।

- गुनेटर पी. (2004). साइना के साथ स्वचालित एन.एम.आर. संरचना की गणना. *मेथड मॉलिक्यूँलर बायोलॉजी*, 278, 353।

- लैंडर ई., वॉटरमैन एम., (1999). जीवन के रहस्य: आणविक जीवविज्ञान के लिए गणितज्ञ का परिचय. नेशनल एकेडमी प्रेस, वाशिंगटन डीसी।

- लेहनिंगर ए., नेल्सन डी., कॉक्स एम, (2005). जैव रसायन का सिद्धांत. चौथा संस्करण, न्यूयॉर्क।

- लेविट एम., चोथिया सी. (1976). ग्लोबुलर प्रोटीन में स्ट्रक्चरल पैटर्न. *नेचर*, 261(5561), 552।

- ओस्फोप्रोपे डी. (2000). एब इनिसियो प्रोटीन फोल्डिग. *करंट ओपिनियन स्ट्रकचरल बायोलॉजी*, 10 (2), 146।

- सिमंस के., कोपरबर्ग सी., हुआंग ई., बेकर, डी. (1997). सिमलेट एंगलिंग और बैज़ियन स्कोर फ़ंक्शन का उपयोग करके समान स्थानीय अनुक्रम के टुकड़ों से प्रोटीन तृतीयक संरचनाओं का सन्योगन. *जे. मोल. बॉय.*, 268 (1), 209।

- ट्रामोन्तानो ए. (2006). प्रोटीन संरचना भविष्यवाणी. संस्करण एक, जॉन विले एंड संस, वेनहेम।

संकाय सदस्य

डॉ. अनिल राय

डॉ. यू. बी. अंगड़ि

डॉ. कृष्ण कुमार चतुर्वेदी

डॉ. शशि भूषण लाल

डॉ. मो. समीर फ़ारूक़ी

डॉ. अनु शर्मा

डॉ. संजीव कुमार

डॉ. मीर आसिफ इकबाल

डॉ. द्विजेश चंद्र मिश्र

डॉ. सुधीर श्रीवास्तव

डॉ. नीरज बुढलाकोटी

श्री उमेश चंद्र बंदूनी

उद्धरण: श्रीवास्तव, एस., फ़ारूकी, एम. एस., चतुर्वेदी, के. के. एवं कौर, एम. (2020). कृषि जैव सूचना में टूल्स और तकनीकियों का अवलोकन, हिन्दी कार्यशाला, ई–संदर्भ संहिता, भा.कृ.अनु.प.–भारतीय कृषि सांख्यिकी अनुसंधान संस्थान, नई दिल्ली