# Bioinformatics prediction of HIV coreceptor usage

Thomas Lengauer, Oliver Sander, Saleta Sierra, Alexander Thielen & Rolf Kaiser

**As sequencing technology and prediction algorithms improve, HIV genotyping and coreceptor usage prediction are likely to play an increasingly important role in guiding patient prognosis and treatment selection.**

Computational analysis of the variations in the HIV-1 genome sequence that correlate with preferential binding to the CCR5 (C-C-motif receptor 5) or CXCR4 (C-X-C motif receptor 4) coreceptors in the host promises to enhance the prediction of disease pathogenesis and enable the optimization of treatment regimes. With new HIV drugs targeting coreceptors entering the market place, resequencing technology rapidly improving in fidelity, efficiency and cost, and prediction algorithms moving beyond considering simple sequence data, bioinformatics approaches promise to transform AIDS treatment strategy and disease management. In this commentary, we provide an up-to-date account of the analysis of viral coreceptor usage with a focus on the utility and potential of bioinformatics methods.

## Preventing HIV cell entry

There is no cure for AIDS, arguably the most devastating killer among human infectious diseases. Because the AIDS-causing pathogen HIV integrates its genome into that of infected host cells, individuals cannot realistically be cleared of virus once infection is established. Thus, anti-HIV drug therapies concentrate on easing symptoms and prolonging life by reducing viral replication. For this purpose, anti-HIV drugs target essential viral proteins in the replication cycle of HIV, mostly the viral reverse transcriptase and protease. Unfortunately, HIV is highly variable and therefore notorious for quickly acquiring resistance against any drug with which it is confronted. Thus, anti-HIV drug regimens have to be monitored and changed from time to time, as resistance is acquired by the viral population evolving inside a patient. Bioinformatics support has proved quite effective in suggesting new drug therapies in this situation[1].

*Thomas Lengauer, Oliver Sander and Alexander Thielen are at the Max Planck Institute for Informatics, Computational Biology and Applied Bioinformatics, Campus E1 4, 66123 Saarbrücken, Germany and Saleta Sierra and Rolf Kaiser are at the Institute for Virology, University of Cologne, Fürst-Pückler-Strasse 56, 50935 Cologne, Germany.*
*e-mail: lengauer@mpi-inf.mpg.de*

Nevertheless, as drugs are being administered, resistance accumulates not only inside a single patient but also in the whole patient population. In fact, we have recently observed a growing number of primary infections with resistant viral strains. This is one of the reasons that the search for new drugs with new modes of action remains critically relevant—even though about two dozen anti-HIV small-molecule drugs are currently available that block viral genome transcription (nucleoside reverse transcriptase inhibitors, nucleotide reverse transcriptase inhibitors and nonnucleoside reverse transcriptase inhibitors) and processing (protease inhibitors).

New developments in anti-HIV therapy encompass several additional routes to blocking the viral replication cycle[2]. For instance, Merck's (Whitehouse Station, NJ, USA) Isentress (raltegravir), a small-molecule inhibitor of the HIV integrase, the molecule that facilitates integration of the viral genome into the human genome, is about to be licensed[3]. In addition, the peptide drug Fuzeon (enfuvirtide/T20), which inhibits HIV cell entry by blocking the capsid protein gp41, jointly developed by Roche (Basel) and Trimeris (Morrisville, NC, USA), was approved by the US Food and Drug Administration (FDA) four years ago.

## Cellular receptors co-opted by HIV

The earliest point of blocking the viral replication cycle is to prevent the virus from entering the host cell. When HIV enters a host cell its surface glycoprotein gp120 attaches to a CD4 receptor. In this process, the virus also needs to bind to a cellular chemokine receptor, termed a coreceptor. Although in cell culture experiments several coreceptors facilitate viral cell entry, only the two coreceptors CCR5 and CXCR4 are relevant *in vivo*[4]. The choice of the coreceptor by the virus is often termed viral tropism. (Actually, tropism means the selection of the target cell, that is, monocytes or lymphocytes. The same term has become common also to denote coreceptor usage.) Although there are a few reported counterexamples, HIV usually requires the CCR5 coreceptor to facilitate primary infection, irrespective of the transmission route and the predominant viral tropism present in the donor[5]. This is also substantiated by the observation that individuals with the homozygous $\Delta$32-mutation, which renders the gene for the CCR5 coreceptor nonfunctional, are almost always resistant to HIV infection[6].

As the infection progresses, the virus evolves and coreceptor usage may change. The switch to CXCR4 usage occurs in about half of the individuals infected with HIV subtype B, which is prevalent in Europe and North America. The switch occurs less often in subtype C, which is prevalent in Africa and Asia; the reasons for this remain unclear. Under drug therapy, consequent switches back and forth between both coreceptors may also occur. The occurrence of CXCR4-using HIV-1 variants is generally indicative of an advanced stage of infection and associated with accelerated disease progression[7,8]. Consequently, one key goal of therapy is to prevent HIV from switching coreceptors from CCR5 to CXCR4.

## Box 1 *In-silico* prediction methods available online

Although several bioinformatics methods for prediction of coreceptor usage have been proposed over the years, only three of them are available as online tools: WetCat, WebPSSM and, from our laboratory, geno2pheno$_{[coreceptor]}$ (see **Table 1**). At this point, all three systems are restricted to using the V3-loop as viral sequence information input. The servers differ with respect to the methods on which the prediction is based, but also in terms of the data sets on which they are trained and the way in which the input has to be supplied.

Some differences are due to the different release dates of the respective software. As time passes, more steps are automated. WetCat, the oldest system, requires data in a restricted input format, including an alignment of the V3-loop(s) to a specific consensus sequence. In contrast, WebPSSM allows unaligned sequences and builds the alignment itself. The system also takes sequence fragments containing amino acids extending beyond the V3-loop. The third system, geno2pheno$_{[coreceptor]}$, detects and aligns the V3-loop from a given sequence automatically. WebPSSM and geno2pheno$_{[coreceptor]}$ have been trained on much larger data sets than WetCat.

There are also differences in the output of the servers. WetCat classifies viral variants into X4 and non-X4, respectively, whereas WebPSSM displays a quantitative value that estimates how likely it is that a virus uses CXCR4. Similarly, geno2pheno$_{[coreceptor]}$ allows selection of a level of specificity that defines how conservative a prediction should be.

In addition to the predictions solely based on the sequence of the V3-loop, the recently updated version of geno2pheno$_{[coreceptor]}$ enables the user to supply certain additional clinical markers, such as $CD4^+$ T-cell counts. The prediction models incorporating this information have been trained on about 1,000 samples of therapy-naive patients[24].

Currently, several antiviral drugs that target cellular receptors for HIV are also under investigation. For example, Tanox's (Houston; now part of Genentech) humanized monoclonal antibody ibalizumab (TNX-355), which blocks the human cellular CD4 receptor for HIV, is currently in phase 2 trials. The two cellular coreceptors CCR5 and CXCR4 are also being targeted[9]. Compounds targeting CXCR4, like AnorMed's (Langley, BC, Canada; now part of Genzyme) small molecules AMD3100 or AMD070, were successfully applied in cell culture experiments, but clinical studies in phases 1 to 3 have ruled out their use in humans, so far, due to the severity of side effects.

The concern about side effects is lessened for CCR5 as this coreceptor does not seem to be essential in humans—Δ32-homozygous individuals have no major apparent deficiencies. Indeed, Pfizer's (New York) small-molecule maraviroc (Selzentry in the United States and Celsentri in the rest of the world) is the first CCR5 blocker to be approved by the FDA and the European Medicine Agency[10]. In addition, Schering-Plough (Kenilworth, NJ, USA) has just launched a phase 3 trial of another small-molecule CCR5 inhibitor, vicriviroc (Sch-417690)[8].

Because coreceptor blockers target human as opposed to viral proteins, it was originally presumed that resistance acquisition by HIV would be slower than for drugs that directly target viral proteins. However, the virus can take several paths to resistance[11]. In addition to switching to the coreceptor CXCR4, the virus can adapt to use CCR5 in the presence of the coreceptor blocker, or preexisting minority variants can emerge that use CXCR4. In fact, viral strains carrying different maraviroc resistance–conferring mutations within gp120 have been observed both in cell culture experiments and in clinical trials in humans[12]. In those clinical trials, about half of the individuals with therapy failure showed a switch of the coreceptor from CCR5 to CXCR4.

As all of the drugs mentioned either are still in development or have only recently entered the marketplace, strains of HIV resistant against them are just emerging. Once multiple drugs against the same target are in use and data on the relevant viral resistances are available, bioinformatics methods (for a review, see ref. 1) can be applied to selecting appropriate therapies.

### Monitoring viral coreceptor usage

Monitoring coreceptor usage is essential in the development and employment of entry inhibitors in testing and therapy. In the late clinical phases of drug testing, evaluating coreceptor usage is necessary because it is necessary to test whether administration of the drug selects for HIV variants that prefer using the CXCR4 coreceptor. The emergence of viral variants that use CXCR4 is considered harmful by many clinicians because it correlates highly with a worsening of patient clinical status and progression to AIDS, although we still do not know what is cause and what effect in this situation. When applying entry inhibitors in therapy, monitoring coreceptor usage will also allow the prediction of drug efficacy. Drugs that target a coreceptor which can be bypassed by HIV will be ineffective, and their use would incur both unnecessary costs and additional risks for the patient.

### Phenotypic determination of viral coreceptor usage

Viral coreceptor usage can be measured *in vitro* by phenotypic assays. In the historically first assays, the pathogenicity of a virus was tested by cultivation of a patient's lymphocytes with donor lymphocytes or with permanent cell lines. The readout of these early assays amounted to the ability of certain viruses to form syncytia, which are visible as giant cells under the light microscope; such viruses were called syncytium inducing. In contrast, other strains did not lead to the production of syncytia, so in this case virus production could be made visible only by serological or PCR techniques[13]. These viruses were called nonsyncytium inducing. Today, we know that the nonsyncytium-inducing phenotype correlates highly with CCR5 usage (and such viruses are thus termed R5 virus), whereas syncytium-inducing viruses use CXCR4 (and are termed X4 virus). The early methods were not very precise, as both coreceptors were offered to the virus and the readout was based on a manual evaluation of a microscopic image. In addition, these methods were very time consuming because few R5 viruses can be

**Table 1 Webservers for prediction of HIV co-receptor usage**

| Name | URL | Method | Updated | Reference |
|---|---|---|---|---|
| geno2pheno$_{[coreceptor]}$ | http://coreceptor.bioinf.mpi-inf.mpg.de | Support vector machines (SVM) | September 2007 | 20 |
| WetCat | http://genomiac2.ucsd.edu:8080/wetcat | Decision trees, SVMs, charge rule | Original version from 2003 | 21 |
| WebPSSM | http://ubik.microbiol.washington.edu/computing/pssm/ | Position-specific scoring matrices (PSSM) | April 2006 | 22 |

grown easily in cell culture and they need more time to replicate.

Because of these drawbacks, current phenotypic determination of coreceptor usage is based on several different recombinant assays[14,15]. These methods allow rapid and reproducible phenotypic analysis in 'indicator' permanent cell lines, which constitutively express CCR5, CXCR4 or both coreceptors. In these assays, HIV gp120 or the whole *env* region from the viral sample (that is, the region encoding the viral surface protein) is amplified by RT-PCR and cloned into a plasmid, which is transfected into cells. Depending on the type of assay, replication-competent viruses or replication-defective pseudoviruses are produced, or viral Env is expressed on the surface of the transfected cells. The recombinant Env (R5 or X4) can bind only to certain indicator cell lines, namely the ones expressing the respective coreceptor. Readout is facilitated by light signals or fluorescence, which are produced by a reporter gene also present in the plasmid.

The concordance between two different phenotypic assays is in the mid-80% range on clinical samples[16]. Experimental assays measuring coreceptor usage still take weeks to return results. These shortcomings make apparent the need for simpler, faster and cheaper *in silico* procedures for analyzing HIV coreceptor usage (for a review, see ref. 17; see **Box 1** and **Table 1**).

## Genotypic methods for predicting coreceptor usage

Genotypic analysis amounts to *in silico* prediction of viral coreceptor usage from viral sequence information, which can be derived with clonal or population-based methods. In the former, individual virus sequences from the patient sample are cloned. In the latter, the entire viral quasispecies, comprising different viral variants within the individual, is amplified and bulk sequenced[18]. Population-based sequencing has been more accessible to clinical practice but can be expected to be replaced increasingly by clonal methods, such as ultra-deep sequencing or single-genome sequencing, as new sequencing technology becomes more widely available.

As in the analysis of viral drug resistance, genotypic predictions of coreceptor usage aim to replace parts of the experimental effort in daily routine. Genotypic methods, which place a large part of the analysis into the computer, are generally more accessible in daily diagnostic practice because genotyping is much easier than phenotyping. Furthermore, they yield results more quickly, are easier to standardize and are also cheaper.

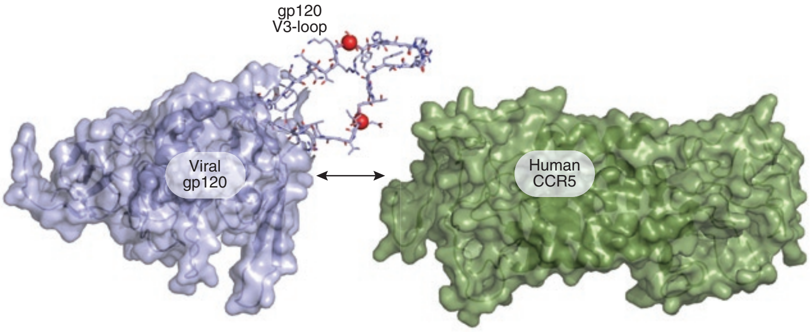

**Figure 1** Structural schema of the proposed molecular interaction between viral gp120 and human CCR5. The viral protein is taken from an X-ray structure[30], the human coreceptor is from a homology model[31]. No measured structures of gp120 in complex with any coreceptor have been published. The orientation of the molecules with respect to each other is derived from modeling studies[31]. The $C_\alpha$ atoms of the two residues at positions 11 and 25 of the V3-loop that are the basis of the 11/25 rule are represented by red balls.

Genotypic methods for predicting coreceptor usage take the relevant parts of the viral genome as input. This is mostly focused on the third highly variable loop (V3-loop) of gp120 (see **Fig. 1**). Even though other parts of the genome seem to influence coreceptor usage as well[19], in the past, a lack of appropriate data has prevented more comprehensive analysis. There are two versions of prediction models: one caters to clonal sequencing, where nonambiguous sequences are subjected to a statistical model; the other processes bulk-sequenced samples including nucleotide mixtures (positions at which two or more different base variants have been detected in the viral strains prevalent in the patient). In both versions, machine-learning methods are used to classify the virus with respect to its coreceptor usage. Different outputs are possible, such as three-way classification into CCR5 users (R5), CXCR4 users (X4) or dual/mixed tropic (R5X4). We find that a dual classification into X4 (allows usage of CXCR4) and non-X4 (does not allow usage of CXCR4) is most useful as it most closely reflects the clinically relevant problem. The methods can output binary as well as quantitative values signifying the confidence of the respective prediction.

Early genotypic methods used very simple rules, such as the 11/25 rule, predicting X4 merely on the basis of the presence of basic side chains in residues 11 or 25 of the V3-loop (see also **Fig. 1**). This rule achieves a sensitivity of around 60% (of the X4 strains predicted as X4) at a specificity of about 92.5% (of the non-X4 strains predicted as non-X4). Recent methods use more complex statistical models, such as support vector machines[20,21] or position-specific scoring matrices (PSSM)[22], which improve the sensitivity by 12 to 17 percentage points at the same level of specificity[23,24]. All of these values have been achieved on clonal-sequenced data. Predictive power decreases substantially when bulk data are used. In the latter case, the 11/25 rule drops to a sensitivity of ~30% at about 93% specificity, whereas the more involved machine-learning methods raise sensitivity to 40–60%[16,24]. The nonuniformity of these values is mainly due to the limited amount of available data, but also reflects the high variance of the phenotypic assays on clinically derived samples. A recently published paper[25] claims that certain phenotypic assays are superior to the genotypic methods developed so far. Such evaluations have to be taken with caution in the face of rapid methodical development. For example, one genotypic system assessed in Low *et al.*[25] has now been replaced by a newer version trained on a larger training set that takes into account additional clinical features, which results in improved prediction power.

More information taken into account by a statistical model affords the possibility of more accurate prediction. So far, only the sequence of the V3-loop has been considered to make the predictions. We have experimented with adding other types of information and found two types to be especially valuable. The first is information on the three-dimensional structure of the V3-loop. Providing the conformation of the V3-loop explicitly to the machine-learning method raises the prediction quality substantially. We have encoded structural information into a descriptor that lists distance distributions between functional groups in the V3-loop. In a comparison on clonal data without insertions or deletions relative to the V3-loop of the crystal structure, we could raise the sensitivity at the specificity level of the 11/25 rule by seven percentage points over that of predictions based on sequence information alone[26]. The second type of additional information to be provided to the statistical model comprises clinical markers, including CD4$^+$ and CD8$^+$ T-cell counts, Δ32-heterozygosity of the patient and information on the variance within the viral quasispecies. This raises the sensitivity from 40% to 63% (with a specificity of 93.5%) on a bulk-sequenced data set of clinically derived patient samples[24].

## Clinical relevance of coreceptor usage prediction

As we mentioned above, in the early phase of HIV infection only non-X4 viruses are usually found. Later in the course of infection, X4 strains are increasingly detected, especially in HIV subtype B. Replication of X4 viruses is not affected by blocking the CCR5 receptor. From the clinical point of view, development of X4 viruses in an individual under treatment with CCR5 blockers can be regarded as a kind of resistance to the drug. The risk of X4 virus emergence rises significantly when the CD4$^+$ T-cell counts drop below the critical threshold of 200 cells/μl. X4 viruses are found to dominate virus population in 10–20% of the patients before onset of first therapy and in 30–60% of the patients that have developed multiple resistances to antiviral drugs during therapy[27,28]. If coreceptor screenings were not applied before therapy was begun, these patients would be administered an ineffective coreceptor blocker. Therefore, time-saving, reliable and widely available methods for genotypic prediction of coreceptor usage are required.

Even in patients with a dominant non-X4 virus, minorities of X4 virus exist, as the recent MOTIVATE (maraviroc plus optimized therapy in viremic antiretroviral treatment experienced patients) study reveals[29]. Furthermore, these minorities can be clinically relevant. In the MOTIVATE study, an X4 minority of 0.1% of total virus in patients administered maraviroc (monotherapy) was observed sufficient for a coreceptor switch within 10 days[29]. It is assumed that the patient's immune system can control these virus strains early in the infection but ceases to be able to do so as the infection progresses. Although the processes underlying this development are not yet understood, it is clear that detection of X4 minorities could be of clinical value.

Here, however, phenotypic methods still have an edge over genotypic methods. The former claim to detect minorities down to ~1%, in contrast to 20% for genotypic methods based on current bulk-sequencing technology. The reason is that the cell culture assays are more sensitive to minorities than the current sequencing methods used for genotyping. Although polymers and enzymes that are used for the currently available sequencing systems are under development, one cannot expect the detection of minorities to become dramatically more sensitive with current bulk-sequencing technology. An alternative with respect to throughput as well as detection limit of minorities is a different sequencing system like the 454 Life Science (Branford, CT, USA) sequencer based on the so-called pyrosequencing approach. With this system, by repeated sequence analysis of the same part of the genome, the detection limit of minorities can be reduced almost at will. The disadvantage of the new system is its high cost; however, as the technology is refined, one can expect the cost of such sequencing to be reduced in the coming years.

Sequence-based genotypic interpretation systems have proved to be very effective for the analysis of resistance versus susceptibility to reverse-transcriptase and protease inhibitors. The present challenge is to use this experience in the development of new tools suitable for coreceptor-usage prediction. Such tools can benefit from the growing availability of ever more detailed data on within-patient virus populations.

## Future perspectives

Genotypic prediction of coreceptor usage is a field of rapidly growing interest. As the first coreceptor antagonist has entered the marketplace and more clinical data become available, we expect a number of innovations to improve the accuracy of genotypic methods. Better high-throughput clonal sequencing technologies will replace bulk sequencing over time. These technologies will not only improve sequencing quality in terms of detection of minorities, but also allow better quantitative handling of quasispecies. In addition, these technologies are likely to produce massive amounts of sequence data on V3-loops and enable sequencing of other regions of the viral envelope protein. Thus, prediction methods should be extended to include non-V3 positions (e.g., the V1- and V2-loops have frequently been proposed to affect coreceptor usage), and optimally the entire HIV *env* gene, to enable elucidation of their role in viral resistance to coreceptor blockers. Larger and better data sets will also assist future research focusing on knowledge of the structural and molecular mechanisms involved in coreceptor usage. These deeper insights into the biology of cell entry can be expected to improve prediction models.

Longer-term objectives will be to uncover the mechanisms of control of X4 virus variants by the immune system early in the course of infection and to predict the evolution of HIV-1 populations in a therapeutic context. The final goal should be to understand the clinical implications of virus heterogeneity, such as how coreceptor usage affects disease progression and therapy selection.

ACKNOWLEDGMENTS
We thank three anonymous reviewers for helpful comments on an earlier draft of this manuscript.

1. Lengauer, T. & Sing, L. *Nat. Rev. Microbiol.* **4**, 790–797 (2006).
2. Opar, A. *Nat. Rev. Drug Discov.* **6**, 258–259 (2007).
3. Cahn, P. & Sued, O. *Lancet* **369**, 1235–1236 (2007).
4. Berger, E.A., Murphy, P.M. & Farber, J.M. *Annu. Rev. Immunol.* **17**, 657–700 (1999).
5. Wolinsky, S.M. *et al. Science* **255**, 1134–1137 (1992).
6. Novembre, J., Galvani, A.P. & Slatkin, M. *PLoS Biol.* **3**, e339 (2005).
7. Regoes, R.R. & Bonhoeffer, S. *Trends Microbiol.* **13**, 269–277 (2005).
8. Westby, M. & van der Ryst, E. *Antivir. Chem. Chemother.* **16**, 339–354 (2005).
9. Rusconi, S. *et al. Curr. Top. Med. Chem.* **7**, 1273–1289 (2007).
10. Dorr, P. *et al. Antimicrob. Agents Chemother.* **49**, 4721–4732 (2005).
11. Moore, J.P. *et al. AIDS Res. Hum. Retroviruses* **20**, 111–126 (2004).
12. Mori, J. *et al. Antivir. Ther.* **12**, S12 (2007).
13. Schuitemaker, H. *et al. J. Virol.* **66**, 1354–1360 (1992).
14. Trouplin, V. *et al. J. Virol.* **75**, 251–259 (2001).
15. Coakley, E., Petropoulos, C.J. & Whitcomb, J.M. *Curr. Opin. Infect. Dis.* **18**, 9–15 (2005).
16. Skrabal, K. *et al. J. Clin. Microbiol.* **45**, 279–284 (2007).
17. Jensen, M.A. & van't Wout, A.B. *AIDS Rev.* **5**, 104–112 (2003).
18. Lehmann, C. *et al. J. Clin. Virol.* **37**, 300–304 (2006).
19. Pastore, C. *et al. J. Virol.* **80**, 750–758 (2006).
20. Sing, T., Beerenwinkel, N. & Lengauer, T. in *Proceedings of the 1st International Workshop on ROC Analysis in Artificial Intelligence (ROCAI 2004)* (eds. Hernández-Orallo, J. *et al.*) 89–96 (IOS Press, Amsterdam, 2004).
21. Pillai, S. *et al. AIDS Res. Hum. Retroviruses* **19**, 145–149 (2003).
22. Jensen, M.A. *et al. J. Virol.* **77**, 13376–13388 (2003).
23. Poveda, E. *et al. AIDS* **21**, 1487–1490 (2007).
24. Sing, T. *et al. Antivir. Ther.* 12, 1097–1106 (2007).
25. Low, A.J. *et al. AIDS* **21**, F17–F24 (2007).
26. Sander, O. *et al. PLoS Comput. Biol.* **3**, e58 (2007).
27. Wilkin, T.J. *et al. Clin. Infect. Dis.* **44**, 591–595 (2007).
28. Brumme, Z.L. *et al. J. Infect. Dis.* **192**, 466–474 (2005).
29. Lewis, M. *et al. Antivir. Ther.* **12**, S65 (2007).
30. Huang, C.C. *et al. Science* **310**, 1025–1028 (2005).
31. Liu, S., Fan, S. & Sun, Z. *J. Mol. Model.* **9**, 329–336 (2003).