NOVAS ABORDAGENS DE ALGORITMOS GENÉTICOS PARA MODELAGEM DE NICHO ECOLÓGICO

F. S. Santana[1]

R. L. Stange[1]; T. C. Giannini[2]; I. Alves - dos - Santos[2]; A. M. Saraiva[1]

[1]Escola Politécnica da USP. Av. Prof. Luciano Gualberto, trav. 3, n. 158. Cidade Universitária, São Paulo, SP. 05508 - 900.
[2]Instituto de Biociências da USP. Rua do Matão, trav. 14, n. 321. Cidade Universitária, São Paulo, SP. 05508 - 900.
fabiana.santana@usp.br

INTRODUÇÃO

O nicho ecológico de uma espécie pode ser descrito como um espaço multidimensional onde as diferentes dimensões representam variáveis distintas, tais como condições ambientais (fatores abióticos) e interações com outras espécies (fatores bióticos) e reflete sua distribuição geográfica (Hutchinson 1957, Brown e Lomolino 2006).

A modelagem de nicho ecológico é uma técnica que combina essas variáveis e projeta um cenário que identifica as regiões potencialmente apropriadas para a ocorrência das espécies, criando um mapa de sua distribuição potencial (Stockwell e Peterson 2002, Anderson *et al.*, 2003, Soberon e Peterson 2005, Soberon 2007). Em geral, a modelagem de nicho ecológico requer o uso de ferramentas computacionais, em função da grande quantidade de dados a serem analisados para a obtenção dos modelos.

A modelagem é composta pelas seguintes etapas (Santana *et al.*, 2008):

1. Definir o experimento de modelagem.
2. Obter os pontos de presença e/ou ausência da(s) espécie(s) e georreferenciar, conferir e corrigir os dados desses pontos.
3. Identificar, adquirir e converter (se necessário) os dados ambientais que serão usados na modelagem (por ex. dados de clima, topografia e vegetação).
4. Selecionar o(s) algoritmo(s) para a geração dos modelos de nicho ecológico.
5. Definir os parâmetros de execução do algoritmo.
6. Gerar o(s) modelo(s) a partir do pacote de software escolhido.
7. Avaliar o modelo para determinar se ele está adequado para a espécie em estudo.
8. Validar o modelo: o conhecimento prévio sobre a espécie em estudo e os resultados estatísticos devem ser considerados pelo pesquisador.

Um dos algoritmos que vem sendo bastante utilizado na modelagem de nicho ecológico é o GARP (*Genetic Algorithm for Rule - set Processing*) (Stockwell e Noble 1992, Stockwell e Peters 1999). Este algoritmo utiliza o conceito de nicho ecológico para definir as regras relacionadas à existência de uma população e, a partir delas, processar novas regras, a fim de realizar a modelagem. Quatro regras são usadas simultaneamente (Stockwell *et al.*, 2006): as regras chamadas Atômicas, Bioclim, Faixas e Logísticas. As regras atômicas usam somente um valor para cada variável na condição da regra, por exemplo, "se a temperatura média é 23°C e a precipitação média é 135mm então a espécie está presente". As regras Bioclim e a regras de faixa, basicamente, baseiam - se nas premissas do programa BIOCLIM (Nix 1986) que produz um "envelope" dos valores ambientais para os quais uma determinada espécie ocorre. Um ponto analisado é predito como presente se estiver contido dentro desse envelope e ausente, em caso contrário. As regras logísticas são uma adaptação dos modelos de regressão logísticos, cuja saída é transformada numa probabilidade.

O uso do GARP na modelagem de nicho ecológico apresenta algumas dificuldades, especialmente a demora para se obter um modelo em alguns casos (e.g.: alta resolução dos dados ou grande quantidade de fatores abióticos) e a impossibilidade de incluir conhecimento *a priori* sobre a biologia das espécies analisadas. Por isso, duas variações do GARP vêm sendo desenvolvidas e testadas: o P - GARP e o AdaptGARP.

O algoritmo P - GARP é a versão paralela do algoritmo GARP para supercomputadores com processamento distribuído, normalmente *clusters* computacionais. O P - GARP quebra a sequência de iterações do GARP, porém mantém as demais funcionalidades que caracterizam um algoritmo genético. Isso aumenta a velocidade do processamento possibilitando a obtenção de modelos em um tempo mais curto ou, no caso de múltiplas análises, a obtenção de mais modelos por período de tempo.

O AdaptGARP é uma versão adaptativa do GARP que utiliza tabelas de decisão adaptativas (Bravo *et al.*, 2007). A partir do uso dessas tabelas é possível definir funções adaptativas que manipulam as regras utilizadas, permitindo assim, alterações dinâmicas no conjunto de regras. Isso define

a adaptatividade da solução. Para a modelagem de nicho ecológico, interessa a inclusão de regras baseadas na biologia das espécies para definir limites e direcionar a aplicação das regras de algoritmos genéticos em geral.

OBJETIVOS

Os objetivos do presente trabalho foram: 1) testar o GARP, P - GARP e AdaptGARP na modelagem de duas espécies e 2) comparar os resultados obtidos. Com isso, é possível ilustrar a aplicação destes novos algoritmos em modelagem de nicho ecológico e fundamentar novas iniciativas para o seu desenvolvimento.

MATERIAL E MÉTODOS

Foram utilizados os pontos de ocorrência de uma espécie de abelha (*Peponapis timberlakei* Hurd & Linsley, 1964- Eucerini, Apidae) e uma planta que é uma das fontes de pólen utilizadas por essa abelha (*Cucurbita palmata* S. Watson - Cucurbitaceae). Ambas ocorrem nos desertos do México e sul dos EUA (Hurd e Linsley 1964). Esses pontos foram obtidos a partir dos seguintes sítios na Internet, que apresentam dados de herbários e coleções entomológicas: USDA (http://plants.usda.gov/), University and Jepson Herbaria (http://ucjeps.berkeley.edu/) e o GBIF (http://www.gbif.org/). Além disso, foi consultado também o artigo de Hurd e Linsley (1964).

A ferramenta computacional utilizada para a modelagem foi o openModeller, cuja descrição está em Santana *et al.*, (2008). A versão do algoritmo GARP utilizada foi a do GARP (*single run*), que é um dos algoritmos disponíveis no openModeller.

Foram utilizadas 37 camadas de variáveis ambientais com grade de 5 minutos de arco de resolução, obtidas no Worldclim (http://www.worldclim.org): temperaturas máximas e mínimas e precipitação para 12 meses e altitude. Os detalhes sobre a base de dados associada a essas camadas podem ser encontrados em Hijmans *et al.*, (2005).

A área sob a curva (AUC - *area under curve*) do gráfico do receptor - operador (ROC - *receiver operating characteristic*) pode ser utilizada como uma medida indicativa da acurácia do modelo (Araújo *et al.*, 2005), e foi calculada a partir dos valores da matriz de confusão (Elith *et al.*, 2006, Philips *et al.*, 2006). Foram feitos dois testes: um teste interno usando a totalidade dos pontos de ocorrência obtidos e um teste externo ou teste independente. Para o teste externo, os dados foram divididos em partições, de forma aleatória e sem reposição, e cada partição foi dividida em dados de teste e dados de treino. Foi usado o limite proposto por Stockwell e Peterson (2002) que estipulam que os dados de treino devem conter no mínimo 20 pontos de ocorrência. A proporção utilizada entre número de pontos de treino e número de pontos de teste foi de 70% e 30% respectivamente, e foram calculados a média e o desvio padrão dos valores de AUC obtidos para cada partição de dados de cada espécie (Araújo *et al.*, 2005). Swets (1988) sugere que os resultados do AUC médio sejam interpretados da seguinte forma: excelente, quando acima de 0,90; bom, entre 0,90 - 0,81; razoável, entre 0,80 - 0,71; pobre, entre 0,70 - 0,61; e falho, entre 0,60 - 0,51. A análise da acurácia do modelo, especialmente baseada nos resultados de AUC, tem sido alvo de muita discussão entre os especialistas. Existe, atualmente, uma extensa bibliografia relacionada com a interpretação desses valores e ainda não há um consenso entre os autores sobre isso (McPherson *et al.*, 2004, Austin 2007, Lobo *et al.*, 2008, Peterson *et al.*, 2008).

Apesar de o AdaptGARP possibilitar a inclusão de regras nas tabelas de decisão, nenhuma regra foi adicionada nos experimentos realizados para este trabalho, uma vez que o intuito era o de comparar o desempenho e os resultados obtidos pelos três algoritmos.

RESULTADOS

A partir da consulta nas bases de dados, foram obtidos 31 pontos de ocorrência para a espécie *P. timberlakei* e 90 para *C. palmata*.

Os mapas obtidos mostram áreas de ocorrência potencial similares para os três algoritmos. As áreas de *P. timberlakei* se estenderam principalmente pelos estados da Califórnia, Arizona e Novo México nos EUA e Sonora no México. Já as de *C. palmata* incluem esses mesmos estados mais Utah e Nevada nos EUA e Baixa Califórnia no México.

Os resultados do AUC obtidos a partir do teste interno, executado com a totalidade de pontos de ocorrência para cada espécie, foram equivalentes:

- *P. timberlakei*: 0,89 (GARP), 0,89 (P - GARP) e 0,89 (AdaptGARP).
- *C. palmata*: 0,85 (GARP), 0,86 (P - GARP) e 0,84 (AdaptGARP).

A espécie *P. timberlakei* apresentou valores de AUC ligeiramente mais elevados que *C. palmata*, no entanto, os resultados dos três algoritmos para as duas espécies localizaram - se dentro da mesma faixa (entre 0,81 e 0,90– considerados bons).

Em relação ao teste externo, realizado com as partições dos dados, foram obtidos os seguintes resultados para o AUC:

- *P. timberlakei*: devido ao número de pontos de ocorrência obtido foi possível dividir os dados dessa espécie em apenas uma partição. Os dados de teste (30% dos dados) resultaram em 0,94 (GARP), 0,94 (P - GARP) e 0,94 (AdaptGARP). Já os dados de treino (70% dos dados) apresentaram 0,88 (GARP), 0,88 (P - GARP) e 0,88 (AdaptGARP).
- *C. palmata*: foram feitas três partições para essa espécie. Os dados de teste (30% dos dados) resultaram em 0,82 ±0,06 (GARP), 0,80 ±0,03 (P - GARP) e 0,82 ±0,06 (AdaptGARP). Já os dados de treino (70% dos dados) apresentaram 0,80 ±0,06 (GARP), 0.78 ±0.06 (P - GARP) e 0,79 ±0,03 (AdaptGARP).

Em relação ao teste externo, a espécie *P. timberlakei* também apresentou valores mais elevados de AUC que a *C. palmata*. Mas, considerando - se os diferentes algoritmos, esses valores foram bem semelhantes, sendo que para *P. timberlakei* eles foram iguais.

Já a duração do tempo de processamento dos modelos utilizando - se todos os pontos de ocorrência (teste interno) foi a seguinte:

- *P. timberlakei*: 1m19,626s (GARP), 1m14,179s (P - GARP) e 1m26,394s (AdaptGARP).

- *C. palmata*: 1m26,824s (GARP), 1m19,865s (P - GARP) e 1m35,994s (AdaptGARP).

Os valores obtidos para o processamento com o PGARP foram os menores.

As diferenças observadas nos valores de AUC no teste interno entre as duas espécies, provavelmente são devidas às características das próprias áreas de ocorrência das espécies analisadas. Apesar de ter sido a espécie com menor número de pontos de ocorrência obtidos, *P. timberlakei* apresentou valores maiores de AUC. No entanto, essa espécie ocorre em uma área menor que *C. palmata*, o que pode ter resultado em um modelo com maior acurácia. Porém, flutuações observadas nos valores de AUC são esperadas e essas diferenças não representam uma variação significativa para algoritmos genéticos. Além disso, o GARP não é um algoritmo determinístico (Lankhosrt, 1996), assim como a maioria dos algoritmos de modelagem, portanto modelos ligeiramente diferentes podem ser obtidos a cada execução.

Em relação ao ganho de tempo do PGARP, foi possível observar que o P - GARP foi mais rápido que os demais algoritmos em todos os testes realizados. Porém, a quantidade de variáveis utilizada para este experimento ainda é considerada pequena, uma vez que em todos os casos, os modelos foram obtidos em menos de 2 minutos. Para avaliar o desempenho do P - GARP neste sentido, ainda são necessários outros experimentos. No entanto, uma vez que o P - GARP altera o processamento das regras realizado pelo GARP original, este experimento é de grande importância para avaliar a qualidade dos modelos gerados. Nesse sentido, os resultados obtidos foram extremamente positivos, pois os modelos gerados pelo P - GARP estão compatíveis com os modelos gerados pelo GARP.

Em relação ao AdaptGARP, é possível, num segundo passo, considerar fatos que sejam do conhecimento do pesquisador, alterando as tabelas de decisão com o objetivo de reduzir a quantidade de dados a serem considerados para as tomadas de decisão do algoritmo. Por exemplo, uma espécie que seja bem conhecida e que só ocorra em regiões desérticas, apresentará certas restrições relativas ao clima, índice pluviométrico e temperatura. Os valores podem ser transformados em valores numéricos de mínimos e máximos, restringindo os extremos da tabela de decisão adaptativa para valores expressos em números reais. Além disso, é possível acrescentar ao ferramental adaptativo, fatores de correlação para o caso de estudos de espécies co - dependentes. Por exemplo, no caso de espécies que dependem de certos polinizadores específicos, é necessário considerar o estudo da distribuição do polinizador na distribuição da espécie vegetal enfocada. Analogamente, o estudo da distribuição do polinizador deve considerar a distribuição da espécie vegetal, uma vez que esta pode ser essencial como fonte de recursos para sua sobrevivência.

CONCLUSÃO

Os três algoritmos - GARP, P - GARP e ADAPTGARP - se mostraram equivalentes em relação às áreas potenciais evidenciadas nos mapas e aos resultados de AUC. As vantagens encontradas foram o melhor desempenho do P - GARP, que pode ser empregado com ganho de tempo em computadores baseados em clusters paralelos, e a possibilidade de se aplicar tabelas de decisões adaptativas através do AdaptGARP. Em trabalhos futuros, a solução aqui apresentada em relação ao AdaptGARP poderá evoluir de forma a permitir inserir nas tabelas de decisão o conhecimento prévio do pesquisador sobre a espécie que está em estudo, com o objetivo de aumentar a qualidade do modelo.

Agradecemos ao Professor João José Neto do Laboratório de Linguagens e Técnicas Adaptativas da Escola de Engenharia Politécnica da USP e à FAPESP (processos 04/15801 - 0 e 04/11012 - 0).

REFERÊNCIAS

Anderson, R. P.; Lew, D. & Peterson, A. T. 2003. Evaluating predictive models of species' distributions: criteria for selecting optimal models. *Ecological Modelling* 162: 211 - 232.

Araujo, M. B.; Pearson, R. G.; Thuiller, W. & Erhard, M. 2005. Validation of species - climate impact models under climate change. *Global Change Biology* 11: 1504 - 1513.

Austin, M. 2007. Species distribution models and ecological theory: A critical assessment and some possible new approaches. *Ecological Modeling* 200: 1 - 19.

Bravo, C.; Neto, J. J.; Santana, F.S. & Saraiva, A.M. 2007. Towards and adaptive implementation of genetic algorithms. *Latin American Workshop on Biodiversity Informatics*. INBI 2007 / CLEI 2007-XXXIII. San José, Costa Rica, p. 9–12.

Brown, J. H. & Lomolino, M. V. 2006. *Biogeografia*. 2a edição. Funpec. Ribeirão Preto. 692p.

Elith, J.; Graham, C. H.; Anderson, R. P.; Dudık, M.; Ferrier, S.; Guisan, A.; Hijmans, R. J.; Huettmann, F., Leathwick, J. R., Lehmann, A., Li, J., Lohmann, L. G., Loiselle, B. A., Manion, G.; Moritz, C.; Nakamura, M.; Nakazawa, Y.; Overton, J. M.; Peterson, A. T.; Phillips, S. J.; Richardson, K. S.; Scachetti - Pereira, R.; Schapire, R. E.; Soberon, J.; Williams, S.; Wisz, M. S. & Zimmermann, N. E. 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography* 29: 129 - 151.

Hijmans, J. R.; Cameron, S. E.; Parra, J. L.; Jones, P. G. & Jarvis, A. 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology* 25: 1965–1978.

Hurd, P. D. Jr. & Linsley, E. G. 1964. The squash and gourd bees-Genera *Peponapis* Robertson and *Xenoglossa* Smith-Inhabiting America North of Mexico (Hymenoptera: Apidae). *Hilgardia* 35: 375 - 477.

Hutchinson, G. E. 1957. Concluding remarks. *Cold Spring Harbour Symposium on Quantitative Biology* 22: 415–427.

Lankhorst, M. M. 1996. Genetic algorithms in data analysis. *Thesis Rijksuniversiteit Groningen*. Printed by: Universiteitsdrukkerij Groningen.

Lobo, J. M.; Jimenez - Valverde, A. & Real, R. 2008. AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography* 17: 145 - 151.

McPherson, J. M.; Jetz, W. & Rogers, D. J. 2004. The effect's of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? *Journal of Applied Ecology* 41: 811 - 823.

Nix, H.A. 1986. A biogeographic analysis of Australian elapid snakes. In: Longmore, R. (Ed.), *Atlas of Elapid Snakes of Australia*. Australian Flora and Fauna. Series Number 7. Australian Government Publishing Service, Canberra, p. 4 - 15.

Peterson, A. T.; Papes, M. & Soberon, J. 2008. Rethinking receiver operating characteristic analysis applications in ecological niche modeling. *Ecological Modelling* 213: 63 - 72.

Phillips, S. J.; Anderson, R. P. & Schapire, R. E. 2006. Maximum entropy modeling of species geographic distributions. *Ecological Modelling* 190: 231–259.

Santana, F. S.; Siqueira, M. F.; Saraiva, A. M. & Correa, P. L. P. 2008. A reference business process for ecological niche modeling. *Ecological Informatics* 3: 75 - 86.

Soberon, J. 2007. Grinnellian and Eltonian niches and geographic distributions of species. *Ecology Letters* 10: 1115–1123.

Soberon, J. & Peterson, A. T. 2005. Interpretation of models of fundamental ecological niches and species' distributional areas. *Biodiversity Informatics* 2: 1 - 10.

Stockwell, D. R. B.; Beach, J. H.; Stewart, A.; Vorontsov, G.; Vieglais, D. & Pereira, R. C. 2006. The use of the GARP genetic algorithm and Internet grid computing in the Lifemapper world atlas of species biodiversity. *Ecological Modelling* 195: 139 - 145.

Stockwell, D. R.B. & Noble, I.R. 1992. Induction of sets of rules from animal distribution data: a robust and informative method of analysis. *Mathematics and Computers in Simulation* 33: 385–390.

Stockwell, D. R.B. & Peters, D. 1999. The GARP modelling system: problems and solutions to automated spatial prediction. *International Journal of Geographical Information Science* 13: 143 - 158.

Stockwell, D. R. B & Peterson, A. T. 2002. Effects of sample size on accuracy of species distribution models. *Ecological Modelling* 148: 1–13.

Swets, K. A. 1988. Measuring the accuracy of diagnostic systems. *Science* 240: 1285–1293.