Methods

Classifying black and white spruce pollen using layered machine learning

Surangi W. Punyasena[1,2], David K. Tcheng[2,3], Cassandra Wesseln[1] and Pietra G. Mueller[4]

[1]Department of Plant Biology, University of Illinois, 505 S. Goodwin Avenue, Urbana, IL, 61801, USA; [2]Illinois Informatics Institute, University of Illinois, 1205 West Clark Street, Urbana, IL, 61801, USA; [3]National Center for Supercomputing Applications, 1205 West Clark Street, Urbana, IL, 61801, USA; [4]Illinois State Museum, Research and Collections Center, 1011 E. Ash Street, Springfield, IL, 62703, USA

Summary

• Pollen is among the most ubiquitous of terrestrial fossils, preserving an extended record of vegetation change. However, this temporal continuity comes with a taxonomic tradeoff. Analytical methods that improve the taxonomic precision of pollen identifications would expand the research questions that could be addressed by pollen, in fields such as paleoecology, paleoclimatology, biostratigraphy, melissopalynology, and forensics.

• We developed a supervised, layered, instance-based machine-learning classification system that uses leave-one-out bias optimization and discriminates among small variations in pollen shape, size, and texture. We tested our system on black and white spruce, two paleoclimatically significant taxa in the North American Quaternary.

• We achieved > 93% grain-to-grain classification accuracies in a series of experiments with both fossil and reference material. More significantly, when applied to Quaternary samples, the learning system was able to replicate the count proportions of a human expert ($R^2 = 0.78$, $P = 0.007$), with one key difference – the machine achieved these ratios by including larger numbers of grains with low-confidence identifications.

• Our results demonstrate the capability of machine-learning systems to solve the most challenging palynological classification problem, the discrimination of congeneric species, extending the capabilities of the pollen analyst and improving the taxonomic resolution of the palynological record.

Introduction

Fossil pollen and spores are used to test hypotheses from a broad cross-section of biological and geological sciences. This extraordinarily rich record has been used to address hypotheses from a diverse array of disciplines: from paleoecological and paleoclimatological investigations across hundreds to millions of years; to the identification of plant speciation and extinction events; to the correlation and biostratigraphic dating of rock sequences; to studies of long-term anthropogenic impacts on plant communities; and to the study of plant–pollinator relationships. The consistency of identifications, from sample to sample and from analyst to analyst, is the foundation of all palynological research. This emphasis on repeatability means that identifications are by necessity restricted to taxonomic categories that can be reliably classified by multiple analysts – most often the genus. However, palynologists have long recognized that species identifications are critical to paleoclimatic and paleoecological interpretations, as generic classifications potentially mask radically different environmental adaptations

(Erdtman, 1931; Cain, 1948; Birks & Birks, 2000). This tension between the consistency of classification and taxonomic resolution is a fundamental and unresolved problem in palynology.

A classic paleoecological example of this tension is the morphological discrimination of black and white spruce pollen (*Picea mariana* and *Picea glauca*, respectively). The problem of species identification in spruce has been a recurring issue in Late Quaternary pollen analysis since the beginnings of North American palynology (Erdtman, 1931; Wilson, 1938; Wilson & Kosanke, 1940; Wilson & Webster, 1942; Cain, 1948). For over eight decades, palynologists have sought to differentiate these species because of their different autecological characteristics; white spruce tends to occupy well-drained upland soils, while black spruce often populates poorly drained lowlands and peatlands (Brubaker *et al.*, 1987). Consequently, the changing abundance of black and white spruce has paleoclimatic significance. The frequency of these two species in North American sediment cores, in addition to their substantial ecological differences, makes *Picea* an important taxon in Late Quaternary paleoclimatic reconstructions and the assess-

ment of regional and global climate models (Bartlein *et al.*, 1998; Jackson *et al.*, 2000).

One promising solution to the challenges of species identification is in the application of image analysis and machine learning to pollen analysis. Computer-based analyses of morphological difference permit palynologists to move beyond the qualitative visual identifications that dominate pollen and spore identifications and allow measurement of the consistency and confidence to classifications. However, previous research applying machine learning to pollen identification (Langford *et al.*, 1990; Li & Flenley, 1999; France *et al.*, 2000; Ronneberger *et al.*, 2002; Li *et al.*, 2004; Treloar *et al.*, 2004; Zhang *et al.*, 2004; Chen *et al.*, 2006; Dell'Anna *et al.*, 2009; Landsmeer *et al.*, 2009; Holt *et al.*, 2011) has largely avoided the hard machine-learning problem of species-level classification, the most challenging classification problem in palynology (one exception is Rodriguez-Damian *et al.* (2006), who focused on the morphological discrimination of three Urticaceae species). Consequently, no previous study has focused on the discrimination of paleoecologically and paleoclimatically significant taxa, where the identification of species is significant to the reconstruction of paleoenvironments.

We extend the known capabilities of machine-based classifications by developing a machine-learning system capable of discriminating between the morphologically similar pollen of black and white spruce. By increasing the range of machine-based identifications to the problem of spruce classification, our results establish that automated learning systems can be used not only to automate the identification of morphologically distinct taxa, as has been shown by previous studies, but also to solve paleoecologically critical classification problems for which there is limited expertise. Our experiments include both modern reference and fossil material, demonstrating that the learning system we have designed is not limited to pristine samples and can work with material that has been altered through taphonomic processes, such as compression or tearing.

Materials and Methods

A layered learning system

Our machine-classification experiments were completed using a nearest-neighbor instance-based supervised layered learning system based on kernel density estimation (named ARLO, Automated Recognition and Layered Optimization). ARLO uses bias optimization to find the most effective combinations of experimental parameters (Tcheng *et al.*, 1989, 1991). At the highest level, bias optimization uses an optimizer to maximize the performance of a learning system by manipulating control parameters of the learning system – the bias space (Fig. 1). Bias space is a formal, parameterized space representing all decisions that determine model performance. Bias space includes, but is not limited to, sample preparation, choice of imaging technology, image resolution, choice of image features, image weighting metrics, choice of training examples, single or set of learning algorithms, and how they are combined. A given point in bias space defines all the variables that control the learning process, specifying both the example representation and the learning algorithm to use.
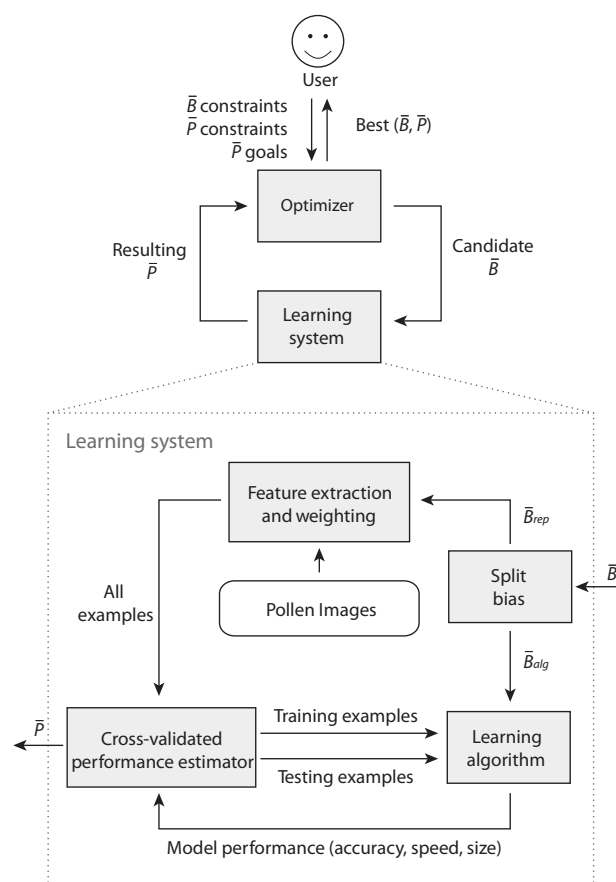

Fig. 1 Schematic of the layered machine-learning classification system. There are two important components to the layered classification system: an optimizer and the learning system. The optimizer seeks to find the bias values (B) that result in the highest performance value (P). The best value discovered within an experimental run is reported as the optimized result. The learning system tries bias values ($\bar{B}$) from the optimizer and outputs a resulting performance values ($\bar{P}$). Within the learning system, there are two areas where bias is optimized: feature representation ($\bar{B}_{rep}$) and learning algorithm ($\bar{B}_{alg}$). Feature representation refers to how the image features are extracted from the pollen images, as described in the Materials and Methods section, and how they are weighted by learning system. Pollen examples are then divided into training and testing sets, which are used to evaluate both the learning algorithm and the feature representation with a cross-validation performance estimator. The performance estimator provides the resulting performance value, $\bar{P}$, to the optimizer.

The optimizer searches bias space by trying a series of candidate points, keeping those that maximize the performance metrics. In this study, we had two distinct performance metrics of interest: grain-to-grain classification accuracy (for both our modern and fossil pollen experiments) and whole slide pollen ratio accuracy (for our fossil pollen experiments). We used a stochastic hill-climbing optimizer to find points in bias space resulting in the best performance. Automating the search in bias space makes the system robust and adaptable to different image recognition problems. Bias search automation also removes reliance on a human experimenter to operate the classification system.

Our learning system is open source and is available for download for use for research purposes from www.sourceforge.net/projects/arlo.

Data sources

We tested our system against two separate data sources: modern reference material derived from vouchered herbarium specimens, where the taxonomic identity of the material was known (Supporting Information, Table S1), and fossil material from Nelson Lake, Illinois, where taxonomic classifications were based on our expert identifications (Table S2). For the modern pollen analysis, in addition to samples of black and white spruce, *P. mariana* (Mill.) Britton, Sterns & Poggenb. and *P. glauca* (Moench) Voss, we included two outgroup genera (*Abies* and *Pinus*) that share a similar morphology, as well as a third North American spruce species, *P. rubens*. For the fossil pollen experiments, we focused our efforts on discriminating the two species that were most abundant in our Nelson Lake samples: black and white spruce. Both sets of pollen material had been prepared following standard protocols (Faegri *et al.*, 1989), with silicone or balsam oil as the mounting medium.

Modern reference pollen samples were from reference collections at the University of Illinois and the Illinois State Museum, isolated from University of Minnesota herbarium specimens (Table S1). Up to 100 grains were imaged from each slide, which represented a single individual tree. An uneven number of representatives were available for each modern class, with an especially limited amount of material available for *P. rubens*. Entire grains, with minimal damage, were chosen for imaging in this analysis of modern material, following the procedure described in the Pollen imaging section.

Fossil spruce examples were from duplicate slides of sediment residues from a published study on Nelson Lake, Illinois (Curry *et al.*, 2007) (Table S2). Ten samples were analyzed, from depths corresponding to high black spruce concentrations, high white spruce concentrations, and roughly equal proportions of the two species. All samples were from fine clay deposits. The preservational quality of these samples were comparable, although more grain damage was observed in the deeper material.

Grains were chosen for imaging for the fossil analysis using a semirandomized method. Student researchers scanned the slides following parallel transects and electronically marked the XYZ location of all saccate grains. Approximately 100 of all the marked grains were then randomly chosen and imaged following the procedure described in the Pollen imaging section. As a result, damaged grains were included, if they were recognizable as a saccate grain by a non-expert. This includes grains that were mechanically or physically changed through tearing, corrosion, or folding.

Each imaged fossil grain was also manually classified by a pollen expert as one of three classes: black spruce, white spruce, or other saccate grain (*Pinus/Abies*) (Table S2); 896 of the 1014 imaged grains were identified as spruce. Identifications were made using the original sample slides, with images as reference to verify that the same grain was being observed. We used a number of morphological features to determine classification manually: grain size, width of saccus at point of attachment, saccus height, angle of saccus attachment, degree of constriction of saccus at point of attachment, saccus shape, endoreticulate pattern of sacci, and relative size of sacci to corpus (Fig. 2; Birks & Peglar, 1980; Hansen & Engstrom, 1985; Lindbladh *et al.*, 2002).

We included a qualitative assessment of expert confidence with each classification in order to record the difficulty of each classification and the certainty of the expert identification. For black and white spruce, these included: 50% (recognized as spruce, but species uncertain), 60% (few key features representative of the species), 70% (several key features representative of the species), 80% (most features representative of the species), 90% (almost all features representative of the species), $\geq$ 95% (all features representative of the species). These numbers capture the self-reported confidence of the human expert for a given classification and so are, by necessity, approximate.

Pollen imaging

We used structured illumination (a Zeiss Apotome fluorescence microscope; Weigel *et al.*, 2009) to produce high-resolution, three-dimensional images. Because of the relatively thin pollen wall of these saccate grains, structured illumination allowed us to capture grain shape and volume in addition to detailed surface images. Images were acquired following a standard manual protocol to minimize variation that would lead to imaging artifacts and potential misdirection of the machine results. Images were taken by multiple researchers, with no
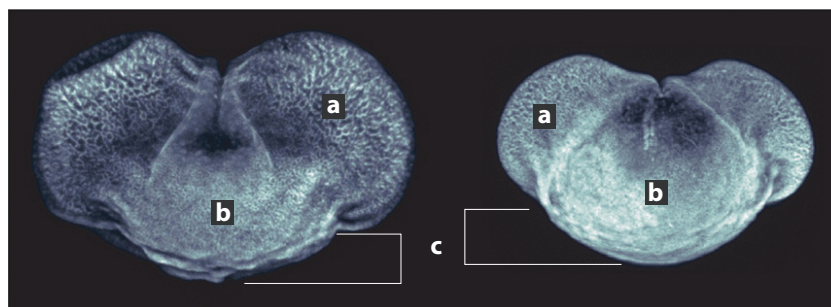

Fig. 2 Morphological features of the *Picea* pollen grain. Illustration of the morphological differences between the pollen of *Picea glauca* (white spruce, left) and *Picea mariana* (black spruce, right). As saccate grains, the two main components of the spruce pollen grain are the saccus (a), a bladder which hydrates and inflates upon contact with the gymnosperm pollen drop; and the pollen body, or corpus (b); c represents the angle of attachment of the saccus to the body. Common features used in the discrimination of black and white spruce include size (*P. glauca* generally having larger grains); angle of the sacci attachment to pollen body (*P. glauca* being more acute); texture of the pollen body (*P. glauca* being finer); the internal reticulate structure of the saccus (*P. glauca* having larger, more circular lumina); and the saccus shape (*P. glauca* having blunter sacci). Note that these traditional measurements of sacci and corpus size and attachment angle can only be taken in the equatorial view (Hansen & Engstrom, 1985; Lindbladh *et al.*, 2002), as illustrated.

one researcher responsible for a single species. Pollen grains were photographed as image stacks using autofluorescence (563 nm excitation frequency (green), 581 nm emission frequency (red)), at 400× magnification (40× EC Plan Neofluor objective, NA 0.75). The shape and depth of the grain were captured as multiple z-focal planes at intervals of half the Nyquist frequency (0.69 μm for this objective; Fig. 3). A typical grain was represented by $c.$ 50 focal slices. Each individual image pixel measured 0.0256 μm^2. Grains were cropped manually, using a bounding box that reached from the maximum width of the grain in the x-axis and the maximum length of the grain in the y-axis. The z-stack was limited to the uppermost and lowermost in-focus planes of the grain.

Example representation and classification

Each image within the z-stack was reduced to a vector of image features. For each new pollen classification problem, the bias optimizer determined the appropriate weights and resolutions for each feature (Fig. 1). These measurements do not directly relate to morphological characters that palynological experts would use, but do describe a large range of morphological variation, totaling > 16 000 dimensions of morphological space. The optimized image features can be categorized into three broad categories:
• *Intensity distribution.* A representation of the probability distribution function of pixel intensity values with a variable number of equally populated bins (quantile values). We used probability distributions with an optimized resolution, from two to 40 quantile bins.
• *Gross shape.* To make gross image comparisons, we compared low-resolution projections of our high-resolution images. These projections captured overall shape and the degree of image coarseness was optimized. We used resolutions as low as 1 × 1 pixels to resolutions as 'high' as 11 × 11 pixels.
• *Texture.* We approximated texture as the change in sign of the first derivative of pixel intensities along a series of horizontal, vertical, or diagonal lines. The line length was optimized, and ranged from 1 to 13 pixels. The texture features were applied to the original high-resolution image as well as down-sampled versions of the image at varying scales.

In addition to these optimized feature groups, we also used three fixed measurements in our classification: image area (the area of an image slice); image aspect (the height-to-width aspect ratio of an image slice); and image depth (the depth of a grain measured as the number of image slices). Grains were classified using a weighted contribution of each image slice. Weighting was by the confidence of image classification, raised to an optimized power. Confidence was measured as the number of nearest neighboring training examples that share the majority classification.

Traditional classification criteria for spruce include a mix of manually assessed quantitative and qualitative characters (Table S3). Researchers consistently use five key features: size (white spruce generally having larger grains); angle of the sacci attachment to pollen body (white spruce being more acute); texture of the pollen body (white spruce being finer); the internal reticulate structure of the saccus (white spruce having larger, more circular lumina); and the saccus shape (white spruce having blunter sacci) (Fig. 2;

Hansen & Engstrom, 1985; Lindbladh *et al.*, 2002). The features we employed in our study coarsely capture these shape, size, and texture characters, but unlike current methods, can be used irrespective of grain orientation.

Learning system evaluation

System accuracy was measured by repeatedly dividing our image data into training and testing sets (Fig. 1). We formed our classification models on a set of training examples and evaluated performance on a separate, randomized set of testing examples. The learning algorithm used the training examples to form the prediction model and applied it to the test examples while measuring its performance in terms of accuracy, speed, and model size. We repeated the experiment with different random partitions and averaged the results. Without bias optimization (i.e. if we had only run a single iteration of the experiment), this accuracy measure would have been an unbiased predictor of future performance. Since we use bias optimization, we know our accuracy is likely optimistic, or overfit. Overfitting refers to the difference in the system's predicted accuracy and its measured accuracy when the system is applied to new data. The greater the optimization, the greater the overfit. We addressed the problem of overfitting in part by using large training samples and through experimental repetition and performance averaging.

Results

Experiments with modern pollen

Our first classification experiments were with modern reference material and involved the identification of three saccate pollen genera (spruce, *Abies*, and *Pinus*) – a three-class problem. The identification of these genera is considered consistent among palynologists, since there are clear differences in grain size, with *Pinus* on average smaller than spruce, and *Abies* on average larger. There are also differences in the relative size of the sacci to pollen body (as defined in Fig. 2), with *Abies* having a smaller ratio of saccus to corpus size, and *Pinus* having the largest. Fir was represented by one species (*Abies balsamea* ($n = 96$)), pine by four species (*Pinus banksiana*, *Pinus strobus*, *Pinus resinosa*, and *Pinus rigida* ($n = 103$)), and spruce by three species (*P. mariana*, *P. glauca*, and *P. rubens*) ($n = 442$; Table S1).

For this modern three-class problem, we achieved 95.2% classification accuracy based on a full 641-fold leave-one-out cross-validation. The features with the most weight in the optimized bias were texture-related (Table S4). All identified *Abies* grains were *Abies*, and 99% of grains identified as *Pinus* were *Pinus*. However, 6% of fir and pine were misclassified as spruce. This higher number of misclassifications was not unexpected; with 4.5 times more spruce grain examples, a given grain was 4.5 times more likely to be classified as spruce than any other taxon.

We next tested the five categories that would be of interest to a palynologist. These classes were the three spruce species, black spruce ($n = 200$), white spruce ($n = 193$), and *P. rubens* ($n = 49$), and our two outgroup genera, *Abies* and *Pinus*. Our final average
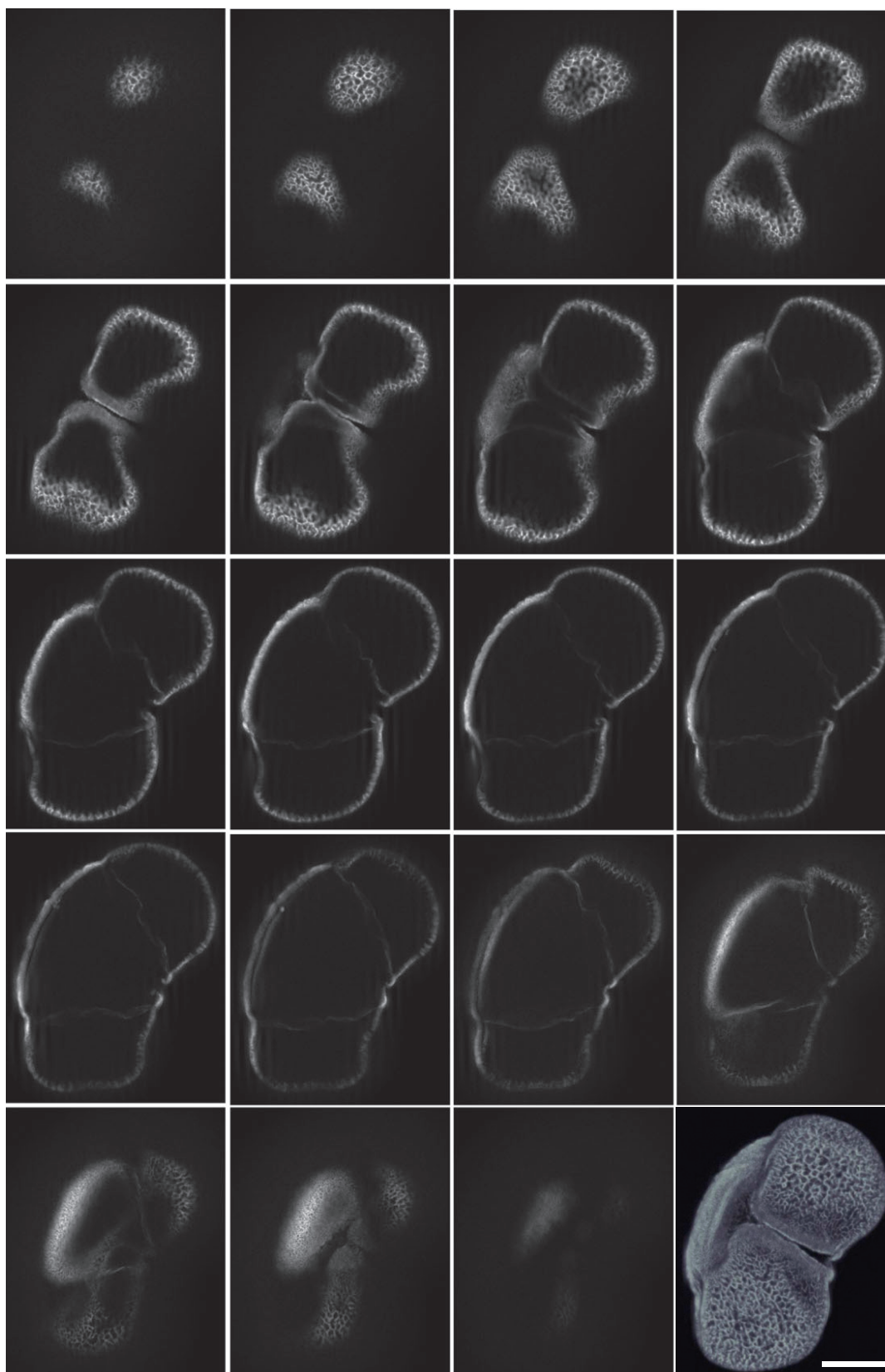
Fig. 3 The pollen image stack. Subsample of images from a single *z*-stack of a *Picea glauca* pollen grain (sample B1500, position 5). A total of 63 images were taken; the slices shown represent approximately every fourth image. Note the ability of structured illumination fluorescence to capture the three-dimensional shape of the grain, including the far pollen wall. The final image (lower right-hand corner) is a maximum intensity projection of the top half of the grain. Each image pixel measures 0.16 μm × 0.16 μm. The original 63 images were taken at 0.69 μm increments in the *z*-plane. Bar, 25 μm.

grain accuracy was 93.3% with a full 641-fold cross-validation (Table S5). Variables with the highest weights were size (measured as image area), followed by texture. *P. rubens* was the most consistently identified of the five classes, with 100% of the grains accurately identified. However, this extremely high classification accuracy suggests that imaging artifacts, potentially resulting from low sample size, may have biased the results.

Smaller white spruce and larger *Pinus* grains were misclassified as black spruce, and larger black spruce and smaller *Abies* grains were misclassified as white spruce. This suggests that, despite overlapping ranges, there is also discriminatory information in size. Finally, the relative importance of texture in the final bias optimization is in agreement with previous work on texture-based machine classifications (Langford *et al.*, 1990; Li & Flenley, 1999; Li *et al.*, 2004; Treloar *et al.*, 2004; Zhang *et al.*, 2004) and suggests that texture, a feature that is often unchanged by moderate taphonomic damage such as compression and tearing, may be an important variable in species discrimination.

Experiments with fossil pollen

Our fossil experiments were with the Quaternary pollen samples from Nelson Lake, Illinois (Curry *et al.*, 2007), corresponding to depths with high black spruce concentrations, high white spruce concentrations, and roughly equal proportions of the two species (Table S2). Ten samples were analyzed, with *c.* 100 saccate grains imaged and manually identified from each sample, as described in the Data sources section. Each identification included an assessment of self-reported expert confidence, ranging from 50% (where only the genus was certain) to $\geq$ 95% (where the species was certain).

We first ran a series of experiments to investigate the relationship between expert confidence and learning system performance. Predictably, when the system was trained and tested on high-quality examples (examples classified with high confidence by the expert), it performed better than when trained on low-quality examples (Fig. 4). There was a clear correlation between human and machine uncertainty, verifying that our human expert assessment of the difficulty of the identification problem meaningfully reflected to true problem difficulty. Poorly preserved, damaged, or ambiguous grains were always challenging – to both the expert and to the system. When spruce grains classified with $\geq$ 95% expert confidence ($n = 264$) were used as the basis for training and testing, our classification accuracy was 94.2% based on full cross-validation of all grains (Table S6).

We next conducted a 'slide-level' leave-one-out cross-validation experiment, where all the pollen grains of a given slide were used for either training or testing, but never both. Accuracy was expected to drop, since unknown grains were never compared with training examples from the same slide. Our final accuracy, however, remained stable at 93.8% (Table S7), demonstrating that the learning system is generalizable and can be applied to new slides drawn from sediment samples with similar taphonomic conditions with very high accuracy. The consistency in our reported accuracy is also our strongest evidence that our learning system is not victim to overly optimistic accuracies based on problem overfitting.
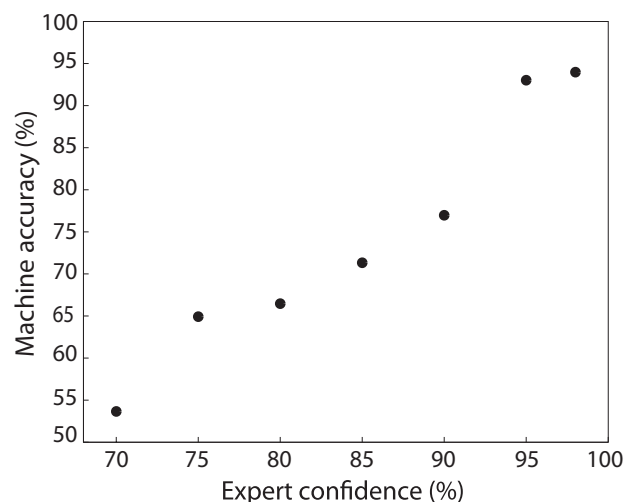

Fig. 4 Relationship between expert confidence and learning system performance. For each expert confidence level (70, 80, 90, 95, and 98%) we created a separate set of training examples that only included examples with that given confidence level. The grains that a human expert could confidently identify were also the grains that our classification system could also most accurately identify, indicating that human assessment of the difficulty of the classification problem related to true problem difficulty.

Reconstructing fossil ratios

Although analysts are trained and tested using grain-to-grain measures of accuracy, consistency of slide-level ratios is the more significant goal. Pollen ratios, as either percentages or pollen densities, must be comparable if palynological data are to be used in any comparative or aggregative study. Our final fossil experiments tested whether we could train the machine classification system to reconstruct slide-level ratios of black and white spruce. We used the difference between human and machine counts, expressed as fractions, as our error metric for optimization and learning.

The final results were striking. We were able to replicate our expert proportions for black and white spruce ($R^2 = 0.78$, $P = 0.007$; Fig. 5) by optimizing the absolute standard error between machine and expert and allowing the system to 'choose' its training examples using bias optimization. Notably, the system chose to use all spruce examples that had been marked with a species classification with an expert confidence of 69% or more (with 50% the lowest possible assigned value). As a result, the learning system's grain-to-grain accuracy fell to 77.5% (Table S8). However, since more grains were included in the final counts by the machine than by the expert, the overall slide-level accuracy was maximized. The high correlation of slide-level ratios was achieved with lower grain-to-grain accuracy, demonstrating that the computer could arrive at similar proportions to human expert counts by using larger numbers of lower-quality data.

Discussion

Morphological differences among congeneric species are often only differences in degree, and the qualitative vocabulary of palynology often cannot convey subtle differences in morphology observed by the analyst. Machine learning circumvents these limitations,
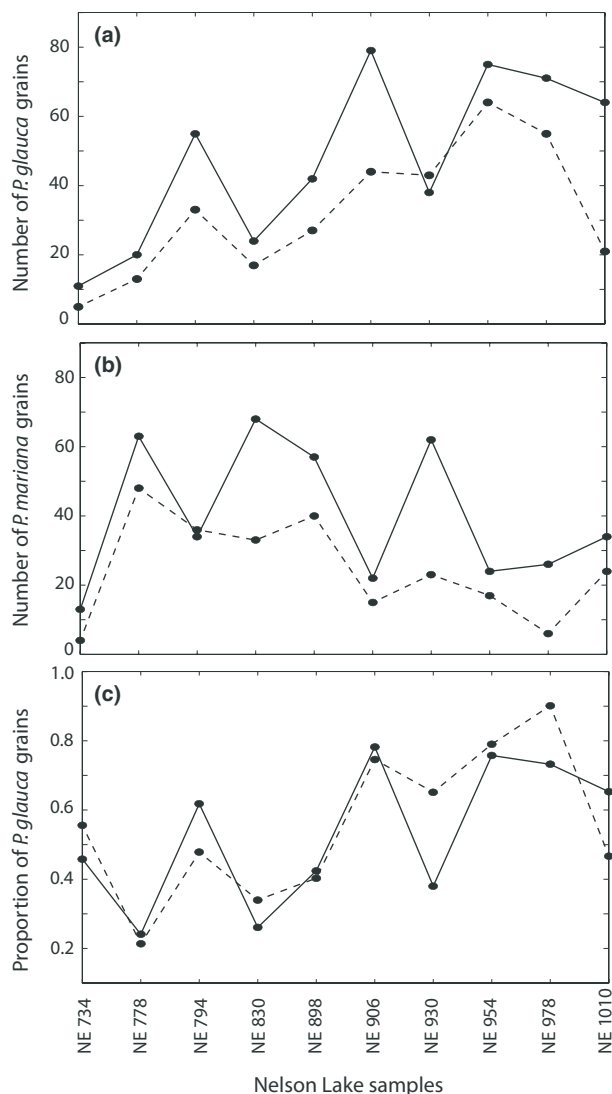
Fig. 5 Comparison of machine and human counts for the fossil Lake Nelson samples. (a, b) Number of grains counted by the classification system (solid line) and number of grains identified with > 80% confidence by the human expert (dashed line) for *Picea glauca* and *Picea mariana*, respectively. (c) Proportion of spruce grains that were identified as *P. glauca*. The machine was able to match human ratios of white to black spruce, based on identifications made with > 80% confidence, by counting as many available grains as possible. Total grains counts per slide are listed in Table S2.

as the quantitative measurements that a computer can make, and the differences a computer can record, are greater than that of the human expert (MacLeod *et al.*, 2010). While the human analyst focuses on recognizing discrete characters, our learning models define a probabilistic feature space. It is the relative placement of an unknown grain in this *n*-dimensional space of morphological variation that is the basis of classification. These measured morphological values do not have to have any biological significance or homology; they need only to define the range of morphological variation observed.

Our machine-based classification system was capable of discriminating the morphologically similar pollen of black and white spruce, from both modern and fossil samples. Our approach is not specific to the features of spruce pollen, so therefore could

potentially be adapted for use with other taxonomic groups. Quantitative machine-based approaches have the additional advantage of objectively quantifying the certainty of identifications, an asset to meta-analyses of palynological data. Since classifications are based on the similarity of an unknown grain to known examples, the strength of this classification is inherently a measure of system certainty. Automation of pollen identification also has uses that extend beyond the North American Quaternary. Besides application to the paleoecological analysis of other locales and other time periods, pollen identification plays a significant role in biostratigraphy, melissopalynology, aerobiology, air quality monitoring, and forensic science. All these fields stand to benefit from the improved ease and accuracy of pollen identification.

The next steps in this research are clear. Comparisons among human experts are needed to place system failure rates in the appropriate context. There is little documentation of consistency in expert identifications of morphologically similar pollen taxa, so the limits of supervised learning systems are still unknown. Additionally, implementing a two-tier cross-validation system that not only optimizes inductive bias but also quantifies the degree of overfitting caused by bias optimization would produce more realistic measurements of system accuracy and results that are less prone to overfitting.

Finally, our results demonstrate that machine-based classifications serve as a means of extending the capabilities of the pollen analyst, improving the taxonomic resolution of the palynological record and delivering the consistent and repeatable species identifications that have largely eluded palynologists for the last century. Machine-based pollen identifications will improve the quality of palynological analysis and, as a result, ultimately expand the climatological and ecological hypotheses that can be addressed by the palynological record. As advances in microscopy and digital imaging increase the speed and affordability of microscope-based data collection, machine classifications and similar image-based technologies will play a crucial role in managing the coming information explosion and transformation of palynological research.

Acknowledgements

References

Bartlein PJ, Anderson KH, Anderson PM, Edwards ME, Mock CJ, Thompson RS, Webb RS, Whitlock C. 1998. Paleoclimate simulations for North America over the past 21,000 years: features of the simulated climate and comparisons with paleoenvironmental data. *Quaternary Science Reviews* 17: 549–585.

Birks HH, Birks HJB. 2000. Future uses of pollen analysis must include plant macrofossils. *Journal of Biogeography* 27: 31–35.

Birks HJB, Peglar SM. 1980. Identification of Picea pollen of Late Quaternary age in Eastern North America: a numerical approach. *Canadian Journal of Botany-Revue Canadienne De Botanique* 58: 2043–2058.

Brubaker LB, Graumlich LJ, Anderson PM. 1987. An evaluation of statistical techniques for discriminating Picea glauca from Picea mariana pollen in Northern Alaska. *Canadian Journal of Botany-Revue Canadienne De Botanique* 65: 899–906.

Cain S. 1948. Palynological studies at Sodon Lake: I. Size-frequency study of fossil spruce pollen. *Science* 108: 115–117.

Chen C, Hendriks EA, Duin RPW, Reiber JHC, Hiemstra PS, de Weger LA, Stoel BC. 2006. Feasibility study on automated recognition of allergenic pollen: grass, birch and mugwort. *Aerobiologia* 22: 275–284.

Curry BB, Grimm EC, Slate JE, Hansen BCS, Konen ME. 2007. The Late-Glacial and Early Holocene geology, paleoecology, and paleohydrology of the Brewster Creek Site, a proposed wetland restoration site, Pratt's Wayne Woods Forest Preserve, and James 'Pate' Philip State Park, Bartlett, Illinois. *Circular 571.* Illinois Department of Natural Resources. Illinois State Geological Survey.

Dell'Anna R, Lazzeri P, Frisanco M, Monti F, Malvezzi Campeggi F, Gottardini E, Bersani M. 2009. Pollen discrimination and classification by Fourier transform infrared (FT-IR) microspectroscopy and machine learning. *Analytical and Bioanalytical Chemistry* 394: 1443–1452.

Erdtman G. 1931. Pollen-statistics: a new research method in paleo-ecology. *Science (New York, NY)* 73: 399–401.

Faegri K, Kaland PE, Krzywinski K. 1989. *Textbook of pollen analysis.* Chichester, UK and New York, NY, USA: Wiley.

France I, Duller AWG, Duller GAT, Lamb HF. 2000. A new approach to automated pollen analysis. *Quaternary Science Reviews* 19: 537–546.

Hansen BCS, Engstrom DR. 1985. A comparison of numerical and qualitative methods of separating pollen of black and white spruce. *Canadian Journal of Botany-Revue Canadienne De Botanique* 63: 2159–2163.

Holt K, Allen G, Hodgson R, Marsland S, Flenley J. 2011. Progress towards an automated trainable pollen location and classifier system for use in the palynology laboratory. *Review of Palaeobotany and Palynology* 167: 175–183.

Jackson ST, Webb RS, Anderson KH, Overpeck JT, Webb T, Williams JW, Hansen BCS. 2000. Vegetation and environment in eastern North America during the Last Glacial Maximum. *Quaternary Science Reviews* 19: 489–508.

Landsmeer SH, Hendriks EA, De Weger LA, Reiber JHC, Stoel BC. 2009. Detection of pollen grains in multifocal optical microscopy images of air samples. *Microscopy Research and Technique* 72: 424–430.

Langford M, Taylor GE, Flenley JR. 1990. Computerized identification of pollen grains by texture analysis. *Review of Palaeobotany and Palynology* 64: 197–203.

Li P, Flenley JR. 1999. Pollen texture identification using neural networks. *Grana* 38: 59–64.

Li P, Treloar WJ, Flenley JR, Empson L. 2004. Towards automation of palynology 2: the use of texture measures and neural network analysis for automated identification of optical images of pollen grains. *Journal of Quaternary Science* 19: 755–762.

Lindbladh MS, O'Connor R, Jacobson GL. 2002. Morphometric analysis of pollen grains for paleoecological studies: classification of Picea from eastern North America. *American Journal of Botany* 89: 1459–1467.

MacLeod N, Benfield M, Culverhouse P. 2010. Time to automate identification. *Nature* 467: 154–155.

Rodriguez-Damian M, Cernadas E, Formella A, Fernandez-Delgado M, De Sa-Otero P. 2006. Automatic detection and classification of grains of pollen based on shape and texture. *IEEE Transactions on Systems Man and Cybernetics Part C-Applications and Reviews* 36: 531–542.

Ronneberger O, Schultz E, Burkhardt H. 2002. Automated pollen recognition using 3D volume images from fluorescence microscopy. *Aerobiologia* 18: 107–115.

Tcheng DK, Lambert BL, Lu SCY, Rendell LA. 1989. Building robust learning systems by combining induction and optimization. Proceedings of the 11th International Joint Conference on Artificial Intelligence 1: 806–812.

Tcheng DK, Lambert BL, Lu SCY, Rendell LA. 1991. AIMS: an adaptive interactive modeling system for supporting engineering decision making. *Machine Learning: Proceedings of the 8th International Workshop* 1: 645–649.

Treloar WJ, Taylor GE, Flenley JR. 2004. Towards automation of palynology 1: analysis of pollen shape and ornamentation using simple geometric measures, derived from scanning electron microscope images. *Journal of Quaternary Science* 19: 745–754.

Weigel A, Schild D, Zeug A. 2009. Resolution in the ApoTome and the confocal laser scanning microscope: comparison. *Journal of Biomedical Optics* 14: 014022.

Wilson LR. 1938. The postglacial history of vegetation in northwestern Wisconsin. *Rhodora* 40: 137–175.

Wilson LR, Kosanke RM. 1940. The microfossils in a pre-Kansan peat deposit near Belle Plaine, Iowa. *Torreya* 40: 1–5.

Wilson LR, Webster RM. 1942. Microfossil studies of three northcentral Wisconsin bogs. *Transcript of the Wisconsin Academy of Science Arts and Letters* 34: 177–193.

Zhang Y, Fountain DW, Hodgson RM, Flenley JR, Gunetileke S. 2004. Towards automation of palynology 3: pollen pattern recognition using Gabor transforms and digital moments. *Journal of Quaternary Science* 19: 763–768.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1 Metadata for the modern reference pollen dataset, taken from collections at the University of Illinois and the Illinois State Museum.

Table S2 Metadata for the Nelson Lake fossil samples, collected from Kane County, Illinois.

Table S3 Morphological characters of *Picea glauca* and *Picea mariana* assessed in three influential studies.

Table S4 Machine-learning outputs for the modern three-class problem, the discrimination of *Picea*, *Abies*, and *Pinus*.

Table S5 Machine-learning outputs for the modern five-class problem, the discrimination of *Picea mariana*, *P. glauca*, *P. rubens*, *Abies*, and *Pinus*.

Table S6 Machine-learning outputs for grain-to-grain analysis of the fossil two-class problem, the discrimination of Nelson Lake *Picea mariana* and *P. glauca*, using grains identified with high expert confidence ($\geq$ 95%).

Table S7 Machine-learning outputs for slide-level analysis of the fossil two-class problem, the discrimination of Nelson Lake *Picea mariana* and *P. glauca*, using grains identified with high expert confidence ($\geq$ 95%).

Table S8 Machine-learning outputs for slide ratio analysis of the fossil two-class problem, the discrimination of Nelson Lake *Picea mariana* and *P. glauca*.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.